CURTIS MILLER

MATH 3070
LECTURE NOTES

Contents

Preface 7

Chapter 1: Overview and Descriptive Statistics 9

Introduction 9
Section 1: Populations, Samples, and Processes 9
Section 2: Pictorial and Tabular Methods in Descriptive Statistics 10
Section 3: Measures of Location 20
Section 4: Measures of Variability 27

Chapter 2: Probability 35

Introduction 35
Section 1: Sample Spaces and Events 35
Section 2: Axioms, Interpretations, and Properties of Probability 41
Section 3: Counting Techniques 50
Section 4: Conditional Probability 59
Section 5: Independence 64

Chapter 3: Discrete Random Variables and Probability Distributions 69

Introduction 69
Section 1: Random Variables 69
Section 2: Probability Distributions for Discrete Random Variables 71
Section 3: Expected Values 81

Section 4: The Binomial Probability Distribution 88
Section 5: Hypergeometric and Negative Binomial Distributions 94
Section 6: The Poisson Probability Distribution 100

Chapter 4: Continuous Random Variables and Probability Distributions 107
Introduction 107
Section 1: Probability Density Functions 107
Section 2: Cumulative Distribution Functions and Expected Values 112
Section 3: The Normal Distribution 119
Section 4: The Exponential and Gamma Distributions 130
Section 5: Other Continuous Distributions 139
Section 6: Probability Plots 149

Chapter 5: Joint Probability Distributions and Random Samples 157
Introduction 157
Section 1: Jointly Distributed Random Variables 157
Section 2: Expected Values, Covariance, and Correlation 166
Section 5: The Distribution of a Linear Combination 175
Section 3: Statistics and Their Distributions 177
Section 4: The Distribution of the Sample Mean 186

Chapter 6: Point Estimation 189
Introduction 189
Section 1: Some General Concepts of Point Estimation 189
Section 2: Methods of Point Estimation 196

Chapter 7: Statistical Intervals Based on a Single Sample 207
Introduction 207
Section 1: Basic Properties of Confidence Intervals 207
Section 2: Large-Sample Confidence Intervals for a Population Mean and Proportion 212
Section 3: Intervals Based on a Normal Population Distribution 216
Section 4: Confidence Intervals for the Variance and Standard Deviation of a Normal Population 222

Chapter 8: Tests of Hypotheses Based on a Single Sample 227

Introduction 227

Section 1: Hypotheses and Test Procedures 227

Section 2: z Tests for Hypotheses about a Population Mean 237

Section 3: The One-Sample t Test 242

Section 4: Tests Concerning a Population Proportion 248

Section 5: Further Aspects of Hypothesis Testing 252

Chapter 9: Inferences Based on Two Samples 265

Introduction 265

Section 1: z Tests and Confidence Intervals for a Difference Between Two Population Means 265

Section 2: The Two-Sample t Test and Confidence Interval 273

Section 3: Analysis of Paired Data 279

Section 4: Inferences Concerning a Difference Between Population Proportions 283

Section 5: Inferences Concerning Two Population Variances 290

Bibliography 297

Preface

These lecture notes were written to accompany Jay Devore's *Probability and Statistics for Engineering and the Sciences (9th ed.)* (Devore, 2015). They are half-filled notes that students are expected to fill out as the instructor lectures and fills out the notes himself on a projected screen for the students to see. This is my preferred lecturing style, as it allows for definitions to be present without needing to waste time writing them down by hand, example problems to be written but not yet solved, and generally improves the flow of the class. Additionally, R code accompanies the mathematical presentation so that students can see how R integrates with the concepts they learn, something that Devore's book does not do. As the class these notes were written for has an accompanying R programming lab, this is a highly useful feature.

These notes do not stand alone and follow tightly to Prof. Devore's book, and I will never release filled-out notes. Additionally, these notes are not intended to be an introduction to R programming; other notes I have written serve that purpose. The R code here is intended to be "real," written to solve problems "the best way possible" rather than in a way students will immediately understand. Devore's book, and some other resource for learning R programming (such as the lab textbook for the course by Verzani (2014)) *must* accompany these notes for them to be of any use. That said, I believe they make a great supplement to Devore's book.

These notes follow Devore's structure exactly and cover Chapters 1 through 9, the chapters covered by the course. Comments are made in the margins, representing asides that are useful or interesting to know (and might even be test or quiz material) yet serve as asides to the main body of information. The notes follow the famous Tufte style; this allows space for the comments and also for plenty of whitespace for note taking and problem solving. There should be plenty of room for students to write.

I hope you find these notes useful.

Curtis Miller

Chapter 1: Overview and Descriptive Statistics

Introduction

This chapter is devoted to basic statistical ideas and statistical summaries. We start with describing what statistics is, does, and what it uses. Next we see graphical and tabular methods for describing distributions. The last two sections discuss measures of location and measures of spread, respectively.

Section 1: Populations, Samples, and Processes

Data is a collection of facts. A **population** is a group of interest. If we collect data for the entire population, we have conducted a **census**. Usually, though, we collect data for a subset of a population, called a **sample**. Our objective is to use the data in the sample to reach conclusions about the population as a whole.

In a sample we have **observations**, individual data points that consist of **variables**, or quantities/characteristics of interest. **Univariate** data records the value of only one variable for each observation. **Multivariate** data records the value of multiple variables for each observation. **Bivariate** data is a special case of multivariate data; there are two variables quantified.

Categorical variables take values from a finite number of possibilities. **Quantitative** variables, however, take numerical values.[1]

Modern statistics depends heavily on probability theory. **Probability** is the field of mathematics that describes the behavior of objects in the presence of uncertainty (which we refer to as randomness). The diagram below illustrates the relationship between probability and statistics with relation to samples and populations.

[1] This may be the simplest dichotomy of types of data. Stevens (1946) classifies data into **nominal**, **ordinal**, **interval**, and **ratio** types, the first two breaking up the "categorical" data type and the second two breaking up the "quantitative" data type. The data types allow for different operations to be defined for different data; ordinal data allows for order relations, interval for addition and subtraction, and ratio allows for division and multiplication.

How we define a population depends heavily on our problem. In **enumerative studies**, the population is a fixed, finite, tangible group that presently exists. In **analytic studies** the population may not presently exist.

Statistics depends crucially on how data is collected in survey-style, observational studies. If data is collected poorly, the results of analysis cannot be trusted.

Below are two approaches for collecting data *correctly*:

- In a **simple random sample (SRS)**, each member of the population of interest is eligible to be randomly selected to be included in the sample. The usual analogy is that each individual in the population is written on a piece of paper and put in a hat; then, slips of paper are randomly chosen in the hat and those individuals are chosen to be in the sample.[2] The statistical methods seen in this class are appropriate for simple random samples *only*.
- In **stratified sampling**, the population is divided into observable **strata**. A SRS is then selected from individuals in each strata.[3]

[2] For example, a candidate for public office may use the registered voter list to randomly select voters in the area the candidate will represent and ask them who they plan to vote for in the upcoming election.

[3] For example, in a national election, an equal number of voters are selected from each state to participate in a poll.

Convenience sampling selects individuals in a way that is not completely random (in the sense that not all individuals from the population are equally likely to be selected, and the procedure is not intentionally stratified). The results of convenience samples cannot be trusted. Statistical descriptions of error account only for error due to randomness, not due to bad sampling procedures.

Section 2: Pictorial and Tabular Methods in Descriptive Statistics

A **distribution** describes what values a variable takes and how frequently it takes them. This section describes techniques for visualizing distributions of univariate data. Visualization is an important first step in a statistical project, as it reveals patterns that are difficult to describe using numbers only, and could suggest what statistical procedures are appropriate.

In statistics, n usually denotes the **sample size**, or the number of observations in the dataset. To denote the values of the dataset's variable, we often use the notation $x_1, x_2, ..., x_n$, where x_i is the ith observation of the dataset. Unless otherwise stated this notation says nothing about the dataset's values. That is, the data is *not* assumed to be ordered.

Stem-and-Leaf Plot

The first visualization of data is a **stem-and-leaf plot**. This plot is constructed using the following steps:

1. Select the number of leading digits to be the **stem** values. The remaining digits are the **leaf** values.
2. Draw a vertical line and list the stem values to the left of this line, in order.
3. Record the leaf of each observation in the row corresponding to its stem value. (Computers often order the leaf values, but when done by hand this is not necessary.)
4. Somewhere in the display, indicate the units of the stem and leaves. (For example, the stems start at the tens place, and the leaves start at the ones place.)

Example 1

The following is a subset of Macdonell's data on height and finger length of criminals imprisoned in England and Wales (Macdonell, 1902). Here I report only the (rounded) heights of the subset.[4]

```
height <- c(5.55, 5.30, 5.63, 5.30, 5.13,
            5.05, 5.38, 5.96, 5.21, 5.38)
```

Use this dataset to construct a stem-and-leaf plot.

[4] Throughout this course I will be including R code that answers the questions I ask. This is so you can see how to do these techniques in R. *You are not expected to understand any of the code at the start of the course!* I *do not* attempt to simplify the code to account for what you have learned so far in the lab. The more you see R code, though, the more familiar and less scary it will become, and I invite you to revisit these lectures at the end of the course and see how much you can understand. Additionally, I hope some of my code will stimulate your curiosity, including the more complicated code.

```r
stem(height, scale = 2)
```

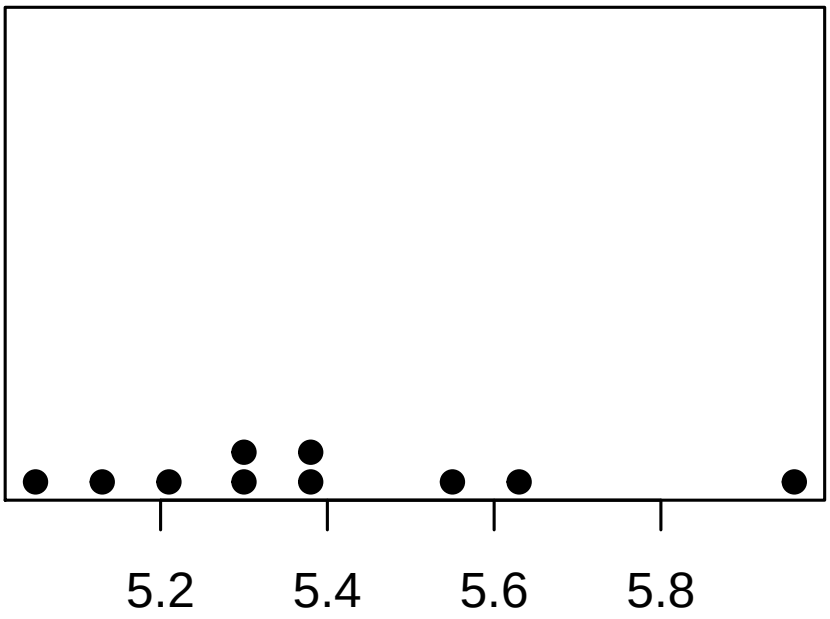
```
##
##   The decimal point is 1 digit(s) to the left of the |
##
##   50 | 5
##   51 | 3
##   52 | 1
##   53 | 0088
##   54 |
##   55 | 5
##   56 | 3
##   57 |
##   58 |
##   59 | 6
```

A **dotplot** represents each data point as a dot along a real number line, putting the point on the line according to its value. If two points would be almost overlapping, they would instead be stacked.

Example 2

Using the data in Example 1, create a dotplot.

```r
stripchart(height, method = "stack", pch = 19, offset = 0.5, at = 0)
```



A quantitative variable is **discrete** if its possible values are countable. It is **continuous** if possible values consist of entire intervals of the real number line (which could be the whole line, in principle).[5]

The **frequency** the value of a variable occurs is the number of times that value was seen in a dataset. For discrete variables it's reasonable to list the frequency of each observed value, but for continuous variables this is not reasonable. Instead, for continuous variables, we list the frequency of a **bin**, which is a range in which a datapoint could be. We would then count how many data points fell within that range.

The **relative frequency** is the frequency a value occured divided by the number of data points. (This is defined analogously for continuous variables.) That is:

[5] As a rule of thumb, discrete variables arise from counting, while continuous variables arise from measurements.

A **frequency distribution** is a tabulation of frequencies or relative frequencies.

Example 3

A statistically minded parent tracks the number of points scored by his daughter's little league soccer team during regular season. Below is the dataset.

```
soccer <- c(9, 6, 5, 5, 5, 6, 2, 8, 3, 4, 8, 1)
```

Construct a frequency distribution for this dataset.

```r
table(soccer)
```

```
## soccer
## 1 2 3 4 5 6 8 9
## 1 1 1 1 3 2 2 1
```

When working with continuous data we need to construct bins when creating a frequency distribution, and list the frequency each bin occurs. How do we do this?

1. Decide on the number of bins. There are rules of thumb for doing this, such as choosing approximately $\sqrt{n}$ bins.[6]
2. Divide the segment of the number line where your data lies into that many equal-length bins.[7]
3. Depending on where each datapoint falls, assign it to a bin. If it falls on a border between bins, assign it to the bin on the right. (In other words, bins are right-inclusive.)
4. Construct a frequency distribution for the bins.

[6] Actually, $n^{1/5}$ may work better.

[7] Some people consider bins of unequal length. When constructing a histogram, do not do this. It makes the histogram more difficult to read correctly.

Example 4

Using the data in Example 1, construct a frequency distribution.

```r
length(height)  # The sample size
```

```
## [1] 10
```

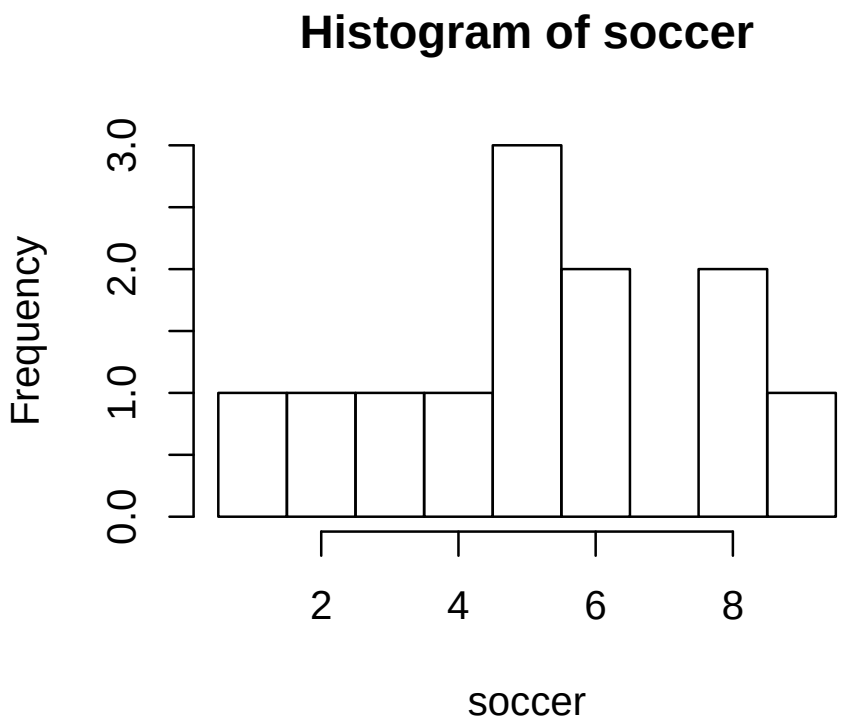
Once we have a frequency distribution, we can construct a **histogram**, a plot for visualizing the distribution of quantitative data. Do the following:

1. Draw a number line and mark the location of the bins. For discrete data, center the bins on the corresponding value.
2. For each class, draw a bar extending from the number line to either the frequency or relative frequency of the number/bin. Do this for each bin.

Example 5

Draw a histogram for the dataset in Example 3 (the soccer dataset).

```r
hist(soccer, breaks = min(soccer):max(soccer + 1) - 0.5)
```



Example 6

For the dataset in Example 1 (the height dataset), create a histogram.

```
hist(height)
```

When looking at plots visualizing distributions we are looking for certain qualities. We want to decide:

- Is the data **unimodal** (only one "peak")? Is it **bimodal** or **multimodal** (multiple "peaks")? Below are illustrations.

- Is the data **positively-skewed**? **Negatively skewed**? **Symmetric**? Below are illustrations.

- Are there **outliers**, points that are distant from the rest of the data?
- How spread out is the data?

A **bar plot** is a method for visualizing categorical (sometimes referred to as **qualitative**) data. To construct a bar plot:

1. List each possible value of the variable and how frequently each value is taken.
2. Draw a horizontal line and along that axis mark each possible value of the variable. The vertical axis will correspond to different possible frequencies.
3. Draw a bar for each category extending to the category's observed frequency.

Example 7

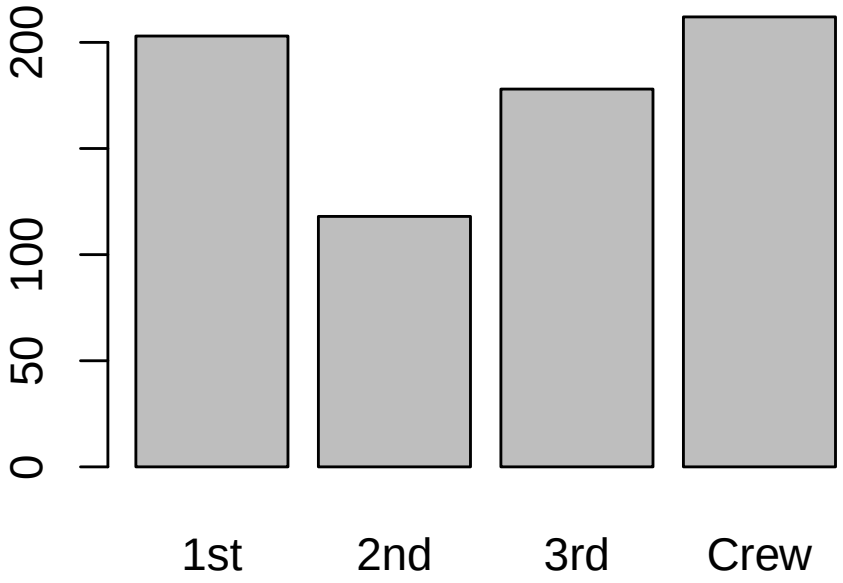
Below is a dataset showing the frequency of the class of passengers aboard the *Titanic* who survived her sinking.

```
(t_survive_class <- apply(Titanic[, , , 2], 1, sum))
```

```
##  1st  2nd  3rd Crew
##  203  118  178  212
```

Create a bar plot for the frequency of each class's survival.

```
barplot(t_survive_class)
```



Section 3: Measures of Location

While visual summaries of data are nice, quantitative summaries are still important for describing datasets. We start with **measures of location**, which tell us where a dataset is located along the number line.

The first and most common measure of location for a sample is the **sample mean**[8], defined for a dataset $x_1, ..., x_n$ below:

[8] There is a physical interpretation of the mean; if you were to construct a dot plot of the data and made that plot a physical object, with a weight for each dot and the number line a teeter-totter, the mean would be the point where the teeter-totter balances.

The **sample proportion** for categorical data is defined below:

Example 8

What is the average number of points your daughter's soccer team scores? (Here's the dataset, as a reminder.)

```
soccer
```

```
##  [1] 9 6 5 5 5 6 2 8 3 4 8 1
```

```
mean(soccer)
```

```
## [1] 5.166667
```

Let's suppose that $r_1, r_2, ..., r_n$ is the *ordered* dataset corresponding to the dataset $x_1, ..., x_n$, so that $r_1 \leq r_2 \leq ... \leq r_n$. The **sample median**[9] is the number that splits this dataset in half. It is defined below:

[9] The physical/geometric interpretation of the median is obvious; when you arrange the data in order, it splits the data in half.

Example 9

Find the median of the first eleven games of your daughter's soccer team. (I have ordered the dataset for you below.)

```
sort(soccer[1:11])
```

```
##  [1] 2 3 4 5 5 5 6 6 8 8 9
```

```
median(soccer[1:11])
```

```
## [1] 5
```

Let $\alpha \in [0,1]$. The $\alpha \times 100$**th percentile** is the number such that roughly $\alpha \times 100\%$ of the data in $r_1, ..., r_n$ lies to the left of that number. Perhaps the most common percentiles are the **quartiles**. The **first quartile** is the 25th percentile, and the **third quartile** is the 75th percentile. The **second quartile** is the median (the 50th percentile).[10]

Here is a procedure for finding quartiles:[11]

1. Find the median of the data $r_1, ..., r_n$.
2. Split the dataset into two datasets at the median. If n is odd, remove the datapoint corresponding to the median.[12]
3. The median of the lower dataset is the first quartile, and the median of the upper dataset is the third quartile.

---

[10] The 0th and 4th quartile are the minimum and maximum of the dataset. All quartiles together form the **five-number summary** of a dataset.

[11] Actually, this is a procedure for finding what your textbook refers to as **fourths**. The difference is negligible so I use the terms interchangeably.

[12] Not everyone does this, so software might give a different answer when computing medians. The difference is usually negligible.

Example 10

Find first and third quartiles for your daughter's first eleven soccer games.

Example 11

Find the 10th and 90th percentiles of the height data. (I have listed the data for you below, in order.)

```
sort(height)
```

```
##  [1] 5.05 5.13 5.21 5.30 5.30 5.38 5.38 5.55
##  [9] 5.63 5.96
```

```
quantile(soccer[1:11], c(.25, .75))
```

```
## 25% 75%
## 4.5 7.0
```

```
quantile(height, c(.1, .9))
```

```
##   10%   90%
## 5.122 5.663
```

The sample mean $\bar{x}$ is **sensitive** to outliers; that is, outliers in the dataset can have a profound effect on the sample mean. On the other hand, the sample median $\tilde{x}$ is **insensitive** to outliers, since outliers almost never alter the value of the sample median.

Example 12

Compute both the sample mean and the sample median when the value of your daughter's 12th soccer game is one of the following:

```
(outlier_game <- c(soccer[12], soccer[12] + 3, soccer[12] * 2, max(soccer) * 2))
```

```
## [1]  1  4  2 18
```

```r
## This loop will compute each of the requested values. I will display the result
## in a table, which is formed when the loop runs.
soccer_tab <- sapply(outlier_game, function(g) {
  dat <- c(soccer[1:11], g)
  return(c(g, median(dat), mean(dat)))
})
soccer_tab <- t(soccer_tab)  # Transpose matrix (I don't want this shape)
## Row/column naming
rownames(soccer_tab) <- 1:nrow(soccer_tab)
colnames(soccer_tab) <- c("Outlier Value", "Median", "Mean")
round(soccer_tab, digits = 2)
```

```
##   Outlier Value Median Mean
## 1             1    5.0 5.17
## 2             4    5.0 5.42
## 3             2    5.0 5.25
## 4            18    5.5 6.58
```

There is in fact a relationship between the mean and median depending on whether the data is negatively-skewed, positively-skewed, or symmetric, illustrated below:

The median is preferred for skewed data while the mean is preferred for symmetric data. (It is better behaved and has great analytic results.)

So far I've discussed only sample means and medians but *population* means, medians, and percentiles are also defined. They have similar properties to their sample analogues.

A compromise between the mean's sensitivity to outliers and the median's ignorance of nearly all of the dataset is the **trimmed mean**, which I denote by $\bar{x}_{\text{tr}(100\alpha)}$. The trimmed mean is the mean of the data when $100\alpha\%$ of the is removed from each end of the dataset.[13]

[13] It may not be possible to remove $100\alpha\%$ of the data *exactly*. You can approximate it with interpolation.

Example 13

Find $\bar{x}_{\text{tr}(10)}$ for the height data.

```r
mean(height, trim = 0.1)
```

```
## [1] 5.36
```

Section 4: Measures of Variability

Consider the following three datasets:

1	2	3
4	2	1
5	5	3
6	6	6
7	7	9
8	10	11

Construct dot plots for each dataset, then compute the mean and median of each dataset.

Now suppose each dataset represented waiting time (in minutes) for the red line train to arrive to take you home. Which dataset would you prefer to see? Why?

The above example illustrates that measures of center are insufficient for describing a dataset. We also want a **measure of variability**, which describes how "spread out" a dataset is.

How can we measure spread? This should be based on **deviations**. The deviation of data point i is $x_i - \bar{x}$.

Compute $\sum_{i=1}^{n}(x_i - \bar{x})$.

This result suggests we should measure variability with something else. The most common measure for variability is the **sample variance** and the **sample standard deviation**, defined below:

The sample standard deviation can roughly be interpreted as the "typical" deviation of a datapoint from the mean.[14]

There are population analogues to both of these quantities: the **population variance**, σ^2, and the **population standard deviation**, $\sigma = \sqrt{\sigma^2}$.

Ideally you should use software or a calculator to compute the sample variance, but in a pinch you can use this handy formula:

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$$

[14] The sample mean is sensitive to outliers. The sample standard deviation is *more* sensitive to outliers than the sample mean.

Example 14

Compute the sample variance and sample standard deviation of the soccer game scores (listed below, as a reminder).

```
soccer
```

```
##  [1] 9 6 5 5 5 6 2 8 3 4 8 1
```

```
length(soccer)
```

```
## [1] 12
```

```
summary(soccer)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.
##   1.000   3.750   5.000   5.167   6.500
##    Max.
##   9.000
```

```r
var(soccer)  # Sample variance
```

```
## [1] 5.969697
```

```r
sd(soccer)  # Sample standard deviation
```

```
## [1] 2.443296
```

Proposition 1. *Let $x_1, ..., x_n$ be a sample and c a constant. Then:*

1. *If $y_i = x_i + c$ for all i, $s_y^2 = s_x^2$*
2. *If $y_i = cx_i$ for all i, $s_y^2 = c^2 s_x^2$ and $s_y = |c| s_x$*

The **fourth spread** (also known as the **inter-quartile range (IQR)**) is the third quartile minus the first quartile; denote this with f_s. This is another measure of dispersion.

Example 15

Compute the fourth spread for the soccer game scores.

---

f_s can be used for outlier detection. We may call an observation that is further than $1.5 f_s$ from its nearest quartile a **mild outlier**, and an observation that is more than $3 f_s$ away from the nearest quartile an **extreme outlier**.

Example 16

Use the fourth spread to detect outliers in soccer game scores. What is the minimum score needed for a data point to be a mild outlier? Extreme outlier?

A **boxplot** is a plot visualizing a dataset. A boxplot is created in the following way:[15]

1. Compute the minimum, maximum, median, first and third quartiles for the dataset.
2. On a number line, draw a box with one end at the first quartile and the other at the third quartile.
3. Within the box, draw a line at the median.
4. Extend a line from one end of the box to the minimum and a line from the other end to the maximum. (These are called **whiskers**.)

Boxplots give both a sense of location and a sense of spread. They're especially useful when placed side-by-side; they then are called **comparative boxplots**.

[15] Often software will not extend the whiskers of box plots to the extrema of samples, instead ending at the largest value that is *not* an outlier. The outliers are then denoted with dots. R, for example, does this by default. While this is more informative it's more difficult to do by hand. The instructions provided here are good enough when not using software.

Example 17

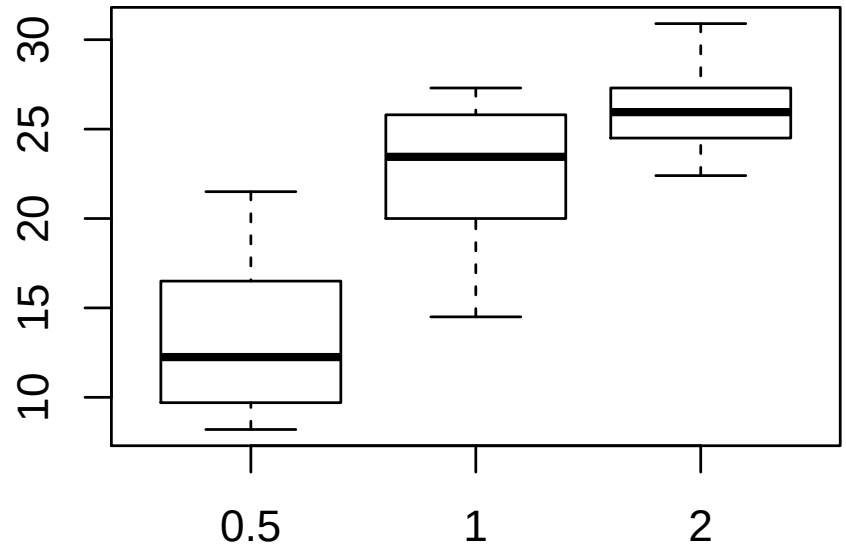
The following dataset contains the tooth growth for guinea pigs given vitamin C via orange juice at three different dosage levels.

```
suppressPackageStartupMessages(library(dplyr))  # Provides %>% operator

OJ <- ToothGrowth %>% filter(supp == "OJ") %>% select(len, dose) %>% unstack %>%
  lapply(sort) %>% as.data.frame
names(OJ) <- c(0.5, 1, 2)
OJ
```

```
##     0.5    1    2
## 1   8.2 14.5 22.4
## 2   9.4 19.7 23.0
## 3   9.7 20.0 24.5
## 4   9.7 21.2 24.8
## 5  10.0 23.3 25.5
## 6  14.5 23.6 26.4
## 7  15.2 25.2 26.4
## 8  16.5 25.8 27.3
## 9  17.6 26.4 29.4
## 10 21.5 27.3 30.9
```

Construct a comparative box plot for the lengths. Compare.

```r
boxplot(len ~ dose, data = ToothGrowth %>% filter(supp == "OJ"))
```



Chapter 2: Probability

Introduction

Next we focus on probability. **Probability** is the mathematical study of randomness and uncertain outcomes. The subject may be as old as calculus. Modern statistics is based on probability theory, often estimating parameters that arise from a probability model. The subject is big and fascinating and sometimes shockingly counterintuitive. In this chapter we introduce basic ideas in probability theory and the theory of counting and combinatorics.

Section 1: Sample Spaces and Events

An **experiment** is an activity or process with an uncertain outcome. Example experiments include:

- Flipping a coin
- Flipping a coin until the coin lands heads-up
- Rolling a six-sided die
- Rolling two six-sided dice
- The time in the morning you wake up

When we have an experiment we need to describe the **sample space**, $\mathcal{S}$[16], which is the set of all possible outcomes of the experiment. A **set** is loosely defined as a collection of objects.[17] **Events** are subsets of the sample space[18], defining possible outcomes of an experiment. The **empty set** or **null event**, $\varnothing$, is a set with no members; it can be thought of as the event that nothing happens.

[16] Another extremely common notation for the sample space is Ω.

[17] This definition cannot be rigorous because it leads to paradoxes. Bertrand Russell was able to find sets that, while legally defined this way, cannot logically exist. Examples include "a set of all sets" and "a set of sets that do not have themselves as members." Axiomatic set theory defines sets in a way that avoids paradoxes but the theory is more complicated than necessary for typical use; the "naive" definition is usually fine.

[18] The sample space is a subset of the sample space and thus is an event, which can be thought of as the event that anything happens.

Example 1

Define a sample space for the experiment of flipping a coin. List all possible events for this experiment.

Example 2

Define a sample space for the experiment of rolling a six-sided die. List three events based on this sample space.

Example 3

Define a sample space describing the experiment of flipping a coin until it lands heads-up. List five events for this sample space.

Example 4

Define a sample space describing the experiment of rolling two six-sided die simultaneoulsy. List three events from this sample space.

Example 5

Define a sample space describing the experiment of waking up in the morning at a particular time, where the time you wake up at (thought of as a real number) is the outcome of interest. List three events from this sample space.

Events can be manipulated in ways to "create" new events. Let A and B be events. The **complement** of A, denoted A'[19] is the set of outcomes of $\mathcal{S}$ not in A, which in words is the event "not A". The **union** of two sets, $A \cup B$, is the set that combines the contents of the sets A and B, which in words means "A or B".[20] The **intersection** of two sets, $A \cap B$, is the set that only includes objects that appear in both A and B, which in words means "A and B".

Two sets are **disjoint** if they have no elements in common. In that case, $A \cap B = \varnothing$.

An intuitive approach to set theory is the use of **Venn diagrams**, where set-theoretic relations are illustrated by depicting objects as points on a plane and denoting set membership with enclosed regions. Below are Venn diagrams illustrating the relations between two sets just described.

[19] Other common notation includes $\overline{A}$ and A^c.

[20] Sets only ever include one copy of each element, so $\{H, H\} = \{H\}$. This implies that if there is a copy of x in both A and B, there will not be two copies of x in $A \cup B$; there is only one copy.

Example 6

Use a Venn diagram to illustrate $(A \cup B)' \cup (A \cap B)$.

Example 7

Consider three sets A, B, and C. Illustrate:

1. $A \cup B \cup C$

2. $A \cap B \cap C$

3. $(A \cap B) \cup (A \cap C) \cup (B \cap C)$

Example 8

Describe the intersection, complement, and union of events described in Examples 1 through 5

Section 2: Axioms, Interpretations, and Properties of Probability

In probability our objective is to assign numbers to events describing how likely that event is to occur. Thus, a **probability measure**, $\mathbb{P}$, is a function taking events as inputs and returning numbers between 0 and 1, and satisfies the following three axioms:

1. $\mathbb{P}(A) \geq 0$
2. $\mathbb{P}(\mathcal{S}) = 1$
3. If $A_1, A_2, ...$ is a sequence of disjoint events (so that for any $i \neq j$, $A_i \cap A_j = \varnothing$), then $\mathbb{P}(A_1 \cup A_2 \cup ...) = \mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$[21]

[21] You may understand this in the more common situation where if $A \cap B = \varnothing$, $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$.

From these, we get all other intuitive relations in probability.

Proposition 2. $\mathbb{P}(\varnothing) = 0$

Proposition 3. $\mathbb{P}(A') = 1 - \mathbb{P}(A)$

Proposition 4. $\mathbb{P}(A) \leq 1$ *for any event* A.

Proposition 5. $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$ *for any events* A *and* B.

Proposition 6. *For any events A, B, and C,*

$$\begin{aligned}\mathbb{P}(A \cup B \cup C) = \mathbb{P}(A) + \mathbb{P}(B) + \mathbb{P}(C) \\ - \mathbb{P}(A \cap B) - \mathbb{P}(A \cap C) - \mathbb{P}(B \cap C) \\ + \mathbb{P}(A \cap B \cap C)\end{aligned}$$

Example 9

Reconsider the experiment of flipping a coin, and assume that the coin is equally likely to land with each face facing up. Assign probabilities to all outcomes in the sample space.

Example 10

Do the same as Example 9, but when rolling a single dice.

Example 11

The dice from Example 10 has been altered with weights. Now, the probability of the dice rolling a 6 is twice as likely as rolling a 1, while all other sides still have the same probability of appearing as before. What is the new probability model?

Example 12

Reconsider the experiment of rolling two six-sided die. It's reasonable to assume that each outcome in $\mathcal{S}$ is equally likely. What, then, is the probability of each outcome in $\mathcal{S}$?

Use this model to find the probability of event E, where:

1. $E = \{\text{At least one dice is a 6}\}$

2. $E = \{\text{The sum of the pips showing on the two die is 5}\}$

3. $E = \{\text{The maximum of the two numbers showing on the die is greater than 2}\}$

Example 13

Reconsider the experiment of flipping a coin until H is seen. What is one way to assign probabilities to all outcomes of this experiment so that we have a legal probability model? Justify your answer.

With this model, answer the following questions:

1. What is the probability the number of flips needed to see the first H exceeds 4?

2. What is the probability the number of flips until the experiment ends is between 3 and 20?

3. What is the probability that an even number of flips is seen before the experiment ends?

Example 14

In a small town, 20% of the population is considered "wealthy", 30% of the population identifies as "black", and 5% of the population is "wealthy" and "black". Select a random individual from this popula-

tion (everyone equally likely to be selected). What is the probability this individual is "wealthy" and "not black"?

What is the probability this individual is neither wealthy nor black?

Example 15

A bag contains balls and blocks. 30% of the bag's contents are balls. An object is either red or blue, and 40% of the objects are red. An object is made of either wood or plastic, and 65% of the objects are wooden. 10% of the objects are wooden balls, 5% of the objects are red balls, and 20% of the objects are red and plastic. 2% of the objects are red plastic blocks.

Reach into the bag and pick out an object at random, each object equally likely to be selected.

1. What is the probability the object selected is a ball, red, or wooden?

2. What is the probability the object is a red wooden ball?

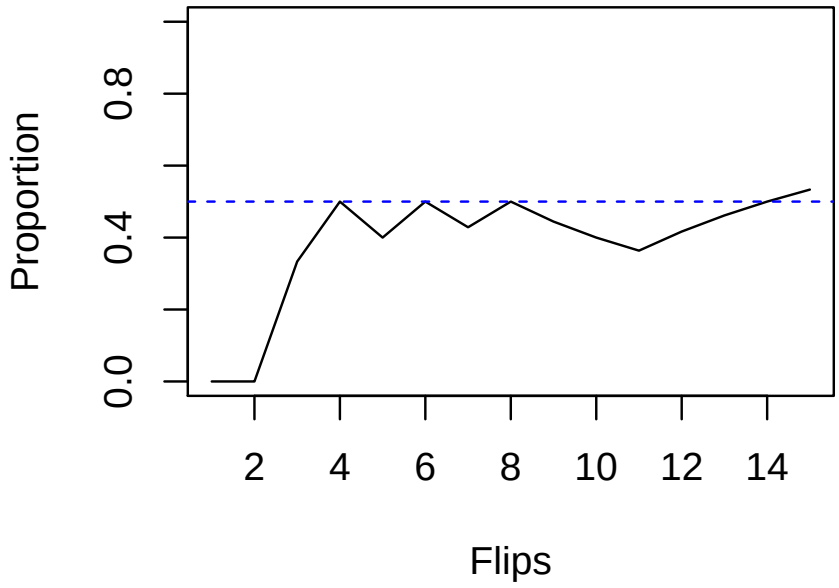
3. What is the probability that the object is a blue plastic block?

How do we interpret probabilies? The **frequentist interpretation of probability**[22] interprets probabilities as the long-run relative frequency as we repeat an experiment many times. For example, if we were to flip a fair coin many times, the proportion of times the coin lands heads up would approach $\frac{1}{2}$.

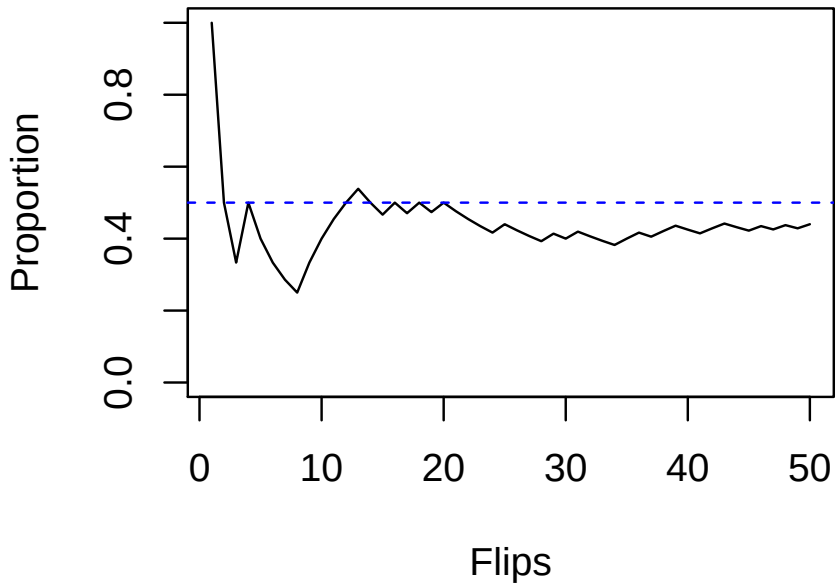
The chart below illustrates this idea.

[22] This interpretation isn't the only one. Any interpretation limits the kind of questions you can obtain probabilities for. In this case, the frequentist interpretation suggests that probabilities can be assigned only to repeatable experiments. While the frequentist interpretation is simple it can lead to convoluted language as we avoid referencing probabilities for nonrepeatable circumstances. The convoluted interpretation of confidence intervals, for example, is due to this interpretation of what a probability means. It turns out though that the rigorous mathematical theory of probability, which is based on measure theory and real analysis, does not care about the "interpretation" of a probability, so all the mathematics remain the same no matter what interpretation we choose.

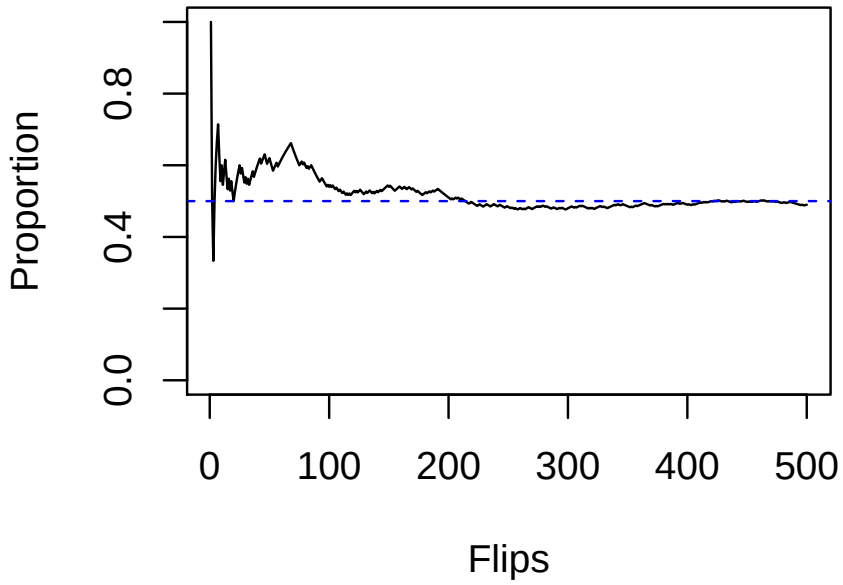
```r
set.seed(11618)    # Choosing a number to set the seed, for replicability
n <- 15
flips <- rbinom(n, 1, 0.5)
heads <- cumsum(flips == 1)
plot(1:n, heads/(1:n), type = "l", ylim = c(0, 1), xlab = "Flips",
     ylab = "Proportion")
abline(h = 0.5, col = "blue", lty = "dashed")
```




```r
n <- 50
flips <- rbinom(n, 1, 0.5)
heads <- cumsum(flips == 1)
plot(1:n, heads/(1:n), type = "l", ylim = c(0, 1), xlab = "Flips",
     ylab = "Proportion")
abline(h = 0.5, col = "blue", lty = "dashed")
```




```r
n <- 500
flips <- rbinom(n, 1, 0.5)
heads <- cumsum(flips == 1)
plot(1:n, heads/(1:n), type = "l", ylim = c(0, 1), xlab = "Flips",
     ylab = "Proportion")
abline(h = 0.5, col = "blue", lty = "dashed")
```



Section 3: Counting Techniques

Consider a burger shop, Bob's Burgers, that offers three types of bread: white, rye, and sourdough. A burger can come with or without cheese. How many burgers are possible?

We first answer this question using a **tree diagram**:

Or we can answer using the **product rule**:

Proposition 7. *If there are n_1 possibilities for choice 1, n_2 possibilities for choice 2, ..., n_k possibilities for choice k, then there are $n_1 n_2 ... n_k = \prod_{i=1}^{k} n_i$ total possible combinations.*

Using the product rule:

Example 16

The sandwich shop Deluxe Deli offers four bread options (white, sourdough, whole wheat and rye), five meat options (turkey, ham, beef, chicken, no meat), six cheese options (cheddar, white cheddar, swiss, American, pepperjack, no cheese), with or whithout lettuce, with or without tomatoes, with or without bacon, with or without mayonaise, and with or without mustard. How many sandwiches are possible?

Suppose that out of n possibilities we will be choosing k. We have two essential questions to answer:

1. Do we choose with or without replacement?
2. Does order matter?

Depending on our answer our question has different solutions, summarized below:

	With replacement	Without replacement
Ordered	n^k	$P_{n,k} = \frac{n!}{(n-k)!}$
Not ordered	$\binom{k+n-1}{n-1}$	$\binom{n}{k} = \frac{n!}{k!(n-k)!}$

Justifications

Example 17

When we roll two six-sided die, we assume each outcome is equally likely (if the dice are different colors). How many possible outcomes are there? What about for three six-sided die?

Example 18

A high school has 27 boys playing men's basketball. In basketball, there are five positions: point guard (PG), shooting guard (SG), small forward (SF), power forward (PF), and center (C). Each assignment of player to position is unique. How many teams can be formed?

Example 19

When playing poker, players draw five cards from a 52-card deck. Every card is distinct, but the order of the draw does not matter. How many hands are possible?

```
## Example 16
6^2
```

```
## [1] 36
```

```
6^3
```

```
## [1] 216
```

```
## Example 17
factorial(27)/factorial(27 - 5)
```

```
## [1] 9687600
```

```
## Example 18
choose(52, 5)
```

```
## [1] 2598960
```

Example 20

You want to choose a dozen donuts from a donut shop. There are eight different kinds of donuts. How many boxes of a dozen donuts are possible?

```
choose(12 + 8 - 1, 8 - 1)
```

```
## [1] 50388
```

For the next few examples, we will be using a standard[23] 52-card deck of playing cards. In this deck, each card belongs to one of four suits: spades (♠), hearts (♡), clubs (♣), and diamonds (♢). Each card has a face value, which is either Ace (A), King (K), Queen (Q), Jack (J), or a number between 2 and 10; there are 13 possible face values. Hearts and diamonds are colored red, while spades and clubs are colored black. The notation 8♢ means "eight of diamonds", K♠ means "king of spades", and so on.

[23] The "standard" deck is the French deck, the most common deck in the English-speaking world. Other European countries have their own traditional decks.

Example 21

A poker hand is "four of a kind" if four cards have the same face value. How many four-of-a-kind hands exist?

Example 22

A poker hand is "full house" of two cards have the same face value and three different cards have another common face value. How many "full house" hands exist?

```
## Example 20
(a1 <- 13 * 48)
```

```
## [1] 624
```

```
## Example 21
(a2 <- 13 * choose(4, 3) * 12 * choose(4, 2))
```

```
## [1] 3744
```

Example 23

A "flush" is a poker hand where all cards belong to the same suit. How many "flush" hands exist (including straight flush hands)?

Example 24

A "straight" is poker hand where the cards can be arranged in sequence: for example, 5♠6♣7♣8♡9♡ is a straight (suit does not matter). A "straight flush" is both a straight and a flush, so it is a flush with all cards belonging to the same suit (and the best possible hand). How many straight flush hands exist? How many straight hands exist (that are *not* straight flushes)?

```
## Example 22
(a3 <- 4 * choose(13, 5))
```

```
## [1] 5148
```

```
## Example 23
(a4 <- 10 * 4^5 - 4 * 10)
```

```
## [1] 10200
```

For finite sample spaces, there is a natural probability measure, defined below for a set A.

Example 25

Use the natural probability measure to compute the probability of each poker hand mentioned in Examples 21 to 24.

```
s <- choose(52, 5)
a1/s  # Exc. 20
```

```
## [1] 0.000240096
```

```
a2/s  # Exc. 21
```

```
## [1] 0.001440576
```

```
a3/s  # Exc. 22
```

```
## [1] 0.001980792
```

```
a4/s  # Exc. 23
```

```
## [1] 0.003924647
```

Section 4: Conditional Probability

Consider flipping a fair coin three times. What is the probability the same face will appear three times?

Now suppose I told you that the first two flips were HH. What is the probability of this event now?

This demonstrates the need for **conditional probability**, which is a probability of an event given the fact that another event has occured. The probability of A given B has occured, denoted $\mathbb{P}(A|B)$, is:

There is an illustration for making this definition intuitive:

Given a conditional probability we can also compute $\mathbb{P}(A \cap B)$:

Given $\mathbb{P}(A|B)$, what is $\mathbb{P}(A'|B)$?

Example 26

Use the definition of conditional probability to compute the probability of the event that all three coins have the same face up when flipped given the first two flips were heads.

Example 27

Suppose that you were dealt two cards of a five-card poker hand, which are K♡8♡. Given this information, what is the probability your complete hand will be a full house?

```
## Hands with KH 8H
(den <- choose(50, 3))
```

```
## [1] 19600
```

```
## Full house hands with KH 8H
(num <- 2 * choose(3, 1) * choose(3, 2))
```

```
## [1] 18
```

```
num / den
```

```
## [1] 0.0009183673
```

```
(num/s) / (den/s)
```

```
## [1] 0.0009183673
```

Example 28

Suppose your five-card poker hand is a flush. What is the probability it is a straight flush?

Example 29

Suppose your five-card poker hand is a straight. What is the probability it is a straight flush?

Example 30

In a certain village 20% of individuals are considered "wealthy" and 35% are considered "black". Among blacks, 60% are not considered "wealthy". If you chose a random individual from this village, what is the probability this individual is "black" and "wealthy"?

A **partition** is a division of $\mathcal{S}$ into sets $A_1, ..., A_n$ such that for $i \neq j$, $A_i \cap A_j = \varnothing$, and $\bigcup_{i=1}^{n} A_i = \mathcal{S}$. Below is an illustration:

Theorem 1 (Law of Total Probability). *Let $A_1, ..., A_n$ be a partition of $\mathcal{S}$ and B be an event. Then:*

$$\mathbb{P}(B) = \sum_{i=1}^{n} \mathbb{P}(B|A_i)\,\mathbb{P}(A_i)$$

We can then state Bayes' Theorem[24]:

Theorem 2 (Bayes' Theorem). *Let $A_1, ..., A_n$ be a partition of $\mathcal{S}$ and B be an event. Then:*

$$\mathbb{P}(A_i|B) = \frac{\mathbb{P}(B|A_i)\,\mathbb{P}(A_i)}{\sum_{j=1}^{n} \mathbb{P}\left(B\middle|A_j\right)\mathbb{P}\left(A_j\right)}$$

[24] Bayes' Theorem is also seen in the simpler form where the partition is A and A', in which case the statement becomes

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\,\mathbb{P}(A)}{\mathbb{P}(B|A)\,\mathbb{P}(A) + \mathbb{P}(B|A')\,\mathbb{P}(A')}$$

Example 31

Roll a fair six-sided dice. Then, after observing the number of pips, roll another dice until more than that number of pips appears. What is the probability that the second die roll will show four pips?

Example 32

In a city Uber and Lyft transport passengers. 95% of drivers work for Uber, and 5% work for Lyft. One day there is a hit-and-run accident and a witness claims that she noticed the driver worked for Lyft. Lyft's defense attorneys subject her to testing, and in testing determine that she correctly identifies a car as belonging to Lyft 90% of the time but will claim a vehicle belongs to Lyft incorrectly 20% of the time. Based on this evidence, how likely is it that the driver who hit the pedestrian worked for Lyft?

Section 5: Independence

Two events A and B are **independent** if $\mathbb{P}(A|B) = \mathbb{P}(A)$. In some sense, information about the event B gives no information about whether A happened.

Use this to compute $\mathbb{P}(B|A)$.

Use this to compute $\mathbb{P}(A'|B)$.

A consequence of this defition of independence:[25]

Below is a graphical representation of independence:[26]

[25] In fact, this may be a more common definition of independence.

[26] Notice that independence is *not* the same as being disjoint. In fact, two disjoint events are not independent except in the most trivial cases. (That is, $\mathcal{S}$ and $\varnothing$ are technically independent.)

Example 33

Consider rolling a 6-sided dice. Show that the events $A = \{\text{Number does not exceed 4}\}$ and $B = \{\text{Number is even}\}$ are independent.

Suppose we have events $A_1, ..., A_n$. These events are **mutually independent** if, for $k \leq n$:

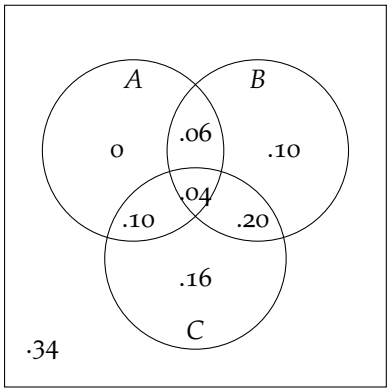
$$\mathbb{P}\left(A_{i_1} \cap A_{i_2} \cap ... \cap A_{i_k}\right) = \mathbb{P}\left(\bigcap_{j=1}^{k} A_{i_j}\right) = \prod_{j=1}^{k} \mathbb{P}\left(A_{i_j}\right) = \mathbb{P}\left(A_{i_1}\right)\mathbb{P}\left(A_{i_2}\right)...\mathbb{P}\left(A_{i_k}\right)$$

This definition cannot be simplified to $\mathbb{P}(A_1 \cap ... \cap A_n) = \mathbb{P}(A_1)...\mathbb{P}(A_n)$, as demonstrated below.[27]

[27] This example was written by George (2004) and is available here: `http://www.engr.mun.ca/~ggeorge/MathGaz04.pdf`

Example 34

Using the diagram below for finding probabilities, compute $\mathbb{P}(A \cap B \cap C)$ and $\mathbb{P}(A)\mathbb{P}(B)\mathbb{P}(C)$. Are A, B, and C mutually independent?



Example 35

We flip eight fair coins. What is the probability of H? TH? TTH? $TTTH$? In general, what is the probability for a sequence of n flips to have $n-1$ T and a H at the end?

Example 36

Below is a system of components. A signal will be sent from one end of the system, and will be successfully transmitted to the other end if no intermediate components fail. Each component functions independently of the others. What is the probability a transmission is sent successfully?

Example 36

Below is another system of components. A signal will be sent from one end of the system, and will be successfully transmitted to the other end if no intermediate components fail. Each component functions independently of the others. What is the probability a transmission is sent successfully?

Chapter 3: Discrete Random Variables and Probability Distributions

Introduction

After we define probability measures and sample spaces, we can talk about random variables. The next two chapters focus on random variables, which translate random outcomes into mathematical objects, such as numbers.[28] This first chapter introduces random variables and a theory for discrete random variables. The second chapter focuses on continuous random variables.

[28] In general random variables can produce any mathematical object. In this class random variables almost always produce real numbers, but in general random variables can also produce vectors or even functions, but this is well beyond the scope of the course.

Section 1: Random Variables

A **random variable** (sometimes abbreviated with **rv**) is a function taking values from the sample space $\mathcal{S}$ and associating numbers with them.[29] Conventional notation for random variables uses capital letters from the end of the English alphabet, while lower-case letters are used to denote a non-random value or outcome. If $\omega \in \mathcal{S}$, the notation $X(\omega) = x$ can be used to say that the value of the random variable X when the outcome ω occurs is x. The set $\{\omega : X(\omega) = x\}$ is the event that an element of $\mathcal{S}$ is drawn that causes the random variable X to equal x, and the set $\{\omega : X(\omega) \in A\}$ is the event that an element of $\mathcal{S}$ is drawn that causes the random variable X to assume a value that is in A.[30] Instead of writing $\mathbb{P}(\{\omega : X(\omega) \in A\})$ we often write $\mathbb{P}(X \in A)$.

[29] From this definition it's clear that random variables are neither random nor variables; they are functions mapping values from $\mathcal{S}$ to some other space, commonly the real numbers $\mathbb{R}$. They can be written $X : \mathcal{S} \to \mathbb{R}$ to emphasize this fact.

[30] The latter set is known as the **preimage** of A under X.

Random variables are commonly classified as being either discrete or continuous.[31] **Discrete** (real-valued) random variables take values in a finite or countably infinite (or enumerable, if you prefer) set with positive probability; these are effectively the only possible values. **Continuous** (real-valued) random variables satisfy the following two properties:

1. The random variable takes values in intervals (possibly infinite

[31] There are random variables considered neither discrete nor continuous. One obvious example is a random variable that is a mixture of discrete and continuous random variables. For example, if a random variable X quantifies the number of hours slept per day, you may have $\mathbb{P}(X = 0) > 0$ and $\mathbb{P}(X = 24) > 0$ but all other outcomes are treated like the continuous case. Yet even then it's possible to define random variables that are not discrete, not continuous, and not a mixture of the two, those these are not seen in practice. In measure-theoretic probability theory, all of these cases are effectively indistinguishable; there isn't a separate theory for each type of random variable. But without this theory we handle discrete and continuous random variables

in length) or disjoint unions of intervals of the real line $\mathbb{R}$ with positive probability.
2. For any $c \in \mathbb{R}$, $\mathbb{P}(X = c) = 0$.

Perhaps the simplest non-trivial random variable is the **Bernoulli random variable**. If X is a Bernoulli random variable, then $\mathbb{P}(X = 1) = 1 - \mathbb{P}(X = 0) = p$, and we say $X \sim \text{Ber}(p)$ to mean this. The set $\mathcal{S}$ on which $X(\omega)$ is defined could be just about anything. An interesting result of probability theory is that if all I gave you was the values of $X(\omega)$ without saying anything about $\mathcal{S}$ or how specifically X assumes values given $\omega \in \mathcal{S}$, it is impossible statistically to determine what $\mathcal{S}$ is. The sample space is effectively forgotten. (In other words, you wouldn't be able to tell the difference between a fair coin and a Bernoulli random variable taking a value of 1 when the coin lands heads-up after being flipped, or a fair die being rolled and a Bernoulli random variable taking a value of 1 when the number rolled is even.)

Example 1

Which of the following random variables are likely to be considered discrete and which continuous? Describe the space of outcomes the random variable takes with positive probability.

1. Flip a coin, record 1 for H, and 0 for T.

2. Roll a die, record the number of pips showing.

3. Roll a die, record 1 for an even number of pips and -1 for an odd number of pips.

4. The time (in minutes) needed to complete a race.

5. The length (in cm) of a hair plucked from a person's head.

6. Roll two dice and record the sum of the number of pips showing.

7. Flip a coin until H is seen and count the number of flips.

Example 2

Consider an experiment of rolling two six-sided die. Define two random variables for this experiment. Are they continuous or discrete?

Section 2: Probability Distributions for Discrete Random Variables

A **probability distribution** for a random variable is a function that describes the probability that a random variable takes on certain values. Discrete rv's are determined completely by the **probability mass function** (abbreviated **pmf**):

The pmf can be visualized using a **line graph**, where a line is placed on each point x of $\mathbb{R}$ that X takes with positive probability and extends to a height representing $p(x)$.

A **probability histogram** functions similarly to a line graph, but is a histogram, with bins centered on x of length 1 (usually) and with height $p(x)$.

Example 3

A fair coin is flipped; $X(H) = 1$ and $X(T) = 0$. Find the pmf of X, $p(x)$. Visualize $p(x)$ with a line graph.

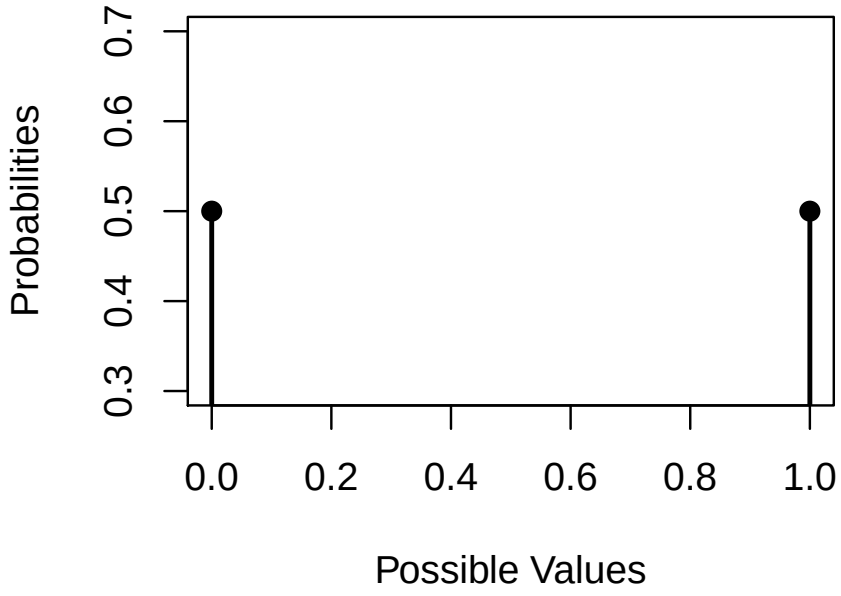
The R package **discreteRV** (Buja et al., 2015) allows for defining and working with discrete random variables in R. (It's pedagogically useful but R's supplied discrete random variable functions are more practical.)

```
suppressPackageStartupMessages(  # Startup messages are annoying
  library(discreteRV)  # An R package for working with discrete random variables
)
```

```
## A statement enclosed in parenthesis prints the variable that was assigned
(X <- RV(0:1, probs = c(1/2, 1/2)))
```

```
## Random variable with 2 outcomes
##
## Outcomes   0   1
## Probs    1/2 1/2
```

```
plot(X)
```



Example 4

Let S be the sum of the number of pips rolled on two dice. Find $p(s)$ and plot it.

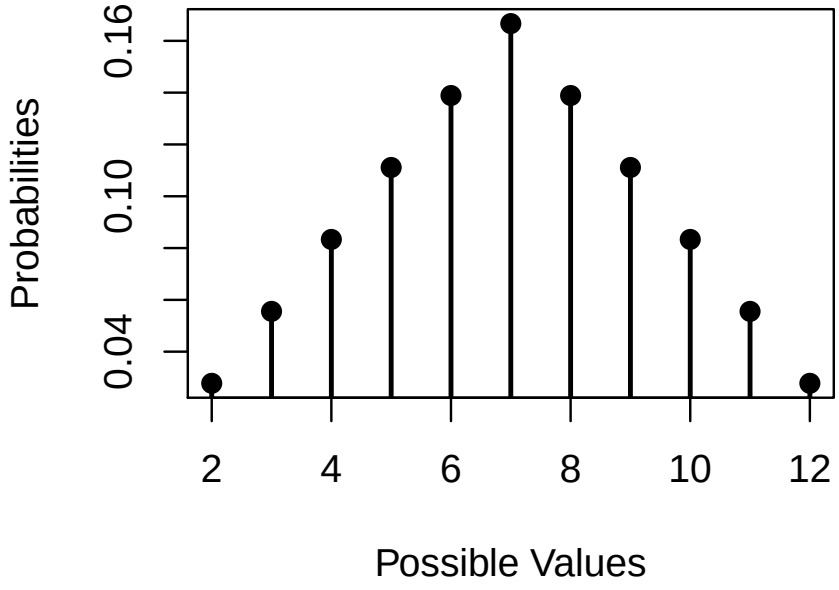
```
(D <- RV(1:6, probs = rep(1/6, times = 6)))
```

```
## Random variable with 6 outcomes
##
## Outcomes   1   2   3   4   5   6
## Probs    1/6 1/6 1/6 1/6 1/6 1/6
```

```
(S <- SofIID(D))
```

```
## Random variable with 11 outcomes
##
## Outcomes    2    3    4    5    6    7    8    9   10   11   12
## Probs    1/36 1/18 1/12  1/9 5/36  1/6 5/36  1/9 1/12 1/18 1/36
```

```
plot(S)
```



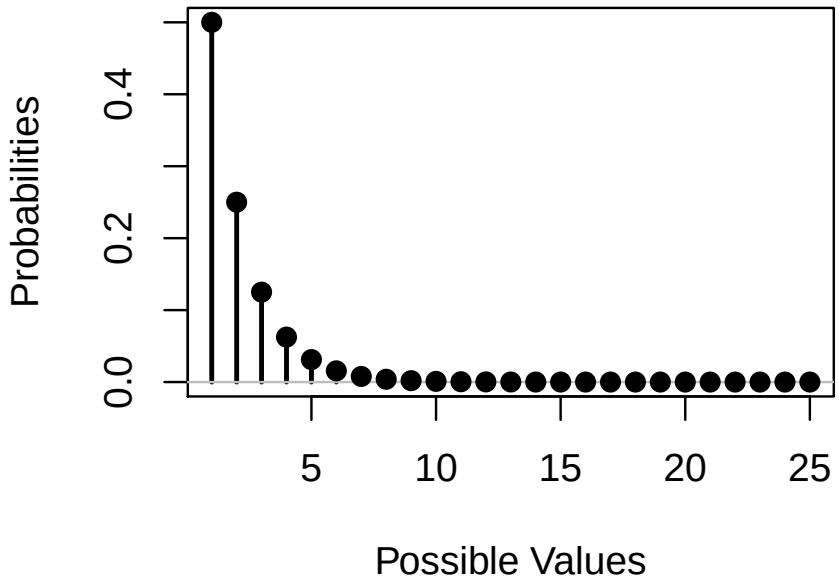
Example 5

Consider flipping a fair coin until H is seen. Let N be the number of flips. Find a pmf describing the distribution of N and plot the first few values of the pmf.

```
(N <- RV("geometric"))
```

```
## Random variable with outcomes from 1 to Inf
##
## Outcomes     1     2     3     4     5     6     7     8     9    10    11    12
## Probs    0.500 0.250 0.125 0.063 0.031 0.016 0.008 0.004 0.002 0.001 0.000 0.000
##
## Displaying first 12 outcomes
```

```
plot(N)
```



Sometimes we describe a distribution in terms of a **parameter**, which is a value that can be set to different possible values to generate a pmf. Probability distributions that differ only in the choice of parameters are called a **family** of distributions.

Example families with parameters include:

Example 6

Confirm that $p(x; N) = \frac{1}{N}$ for $x \in \{1, 2, ..., N\} = [N]$ is a valid pmf. This is the pmf of the discrete uniform distribution, $X \sim \text{DUNIF}(1, N)$.

Example 7

Confirm that $f(n; p) = p(1-p)^{n-1}$ is a valid pmf. This is the pmf of the geometric distribution, $X \sim \text{GEOM}(p)$.

The **cumulative distribution function** (abbreviated **cdf**) is defined below:

Notice the following relation between the cdf and the pdf:

In general, a function $F(x)$ is a cdf if it satisfies the following three properties:[32]

[32] If a function $F(x)$ satisfies these properties than there exists a random variable X with a cdf identical to $F(x)$.

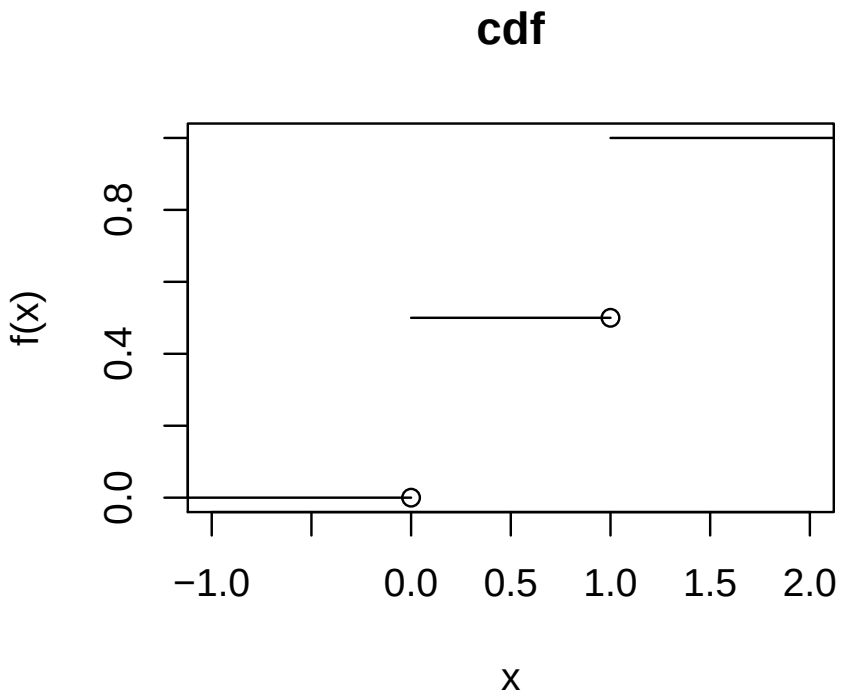
For discrete rv's, cdf's are jump functions, resembling the following plot:

Like a pmf, a cdf completely characterizes a random variable. We can use it for computing probabilities of regions using the following rule:

Example 8

Compute and plot the cdf for a random variable $X \sim \text{Ber}(p)$.

```r
## For p = 1/2

bercdf <- function(q) {pbinom(q, size = 1, prob = 1/2)}
plot(stepfun(0:1, bercdf((-1):1), right = TRUE), verticals = FALSE,
     main = "cdf")
```



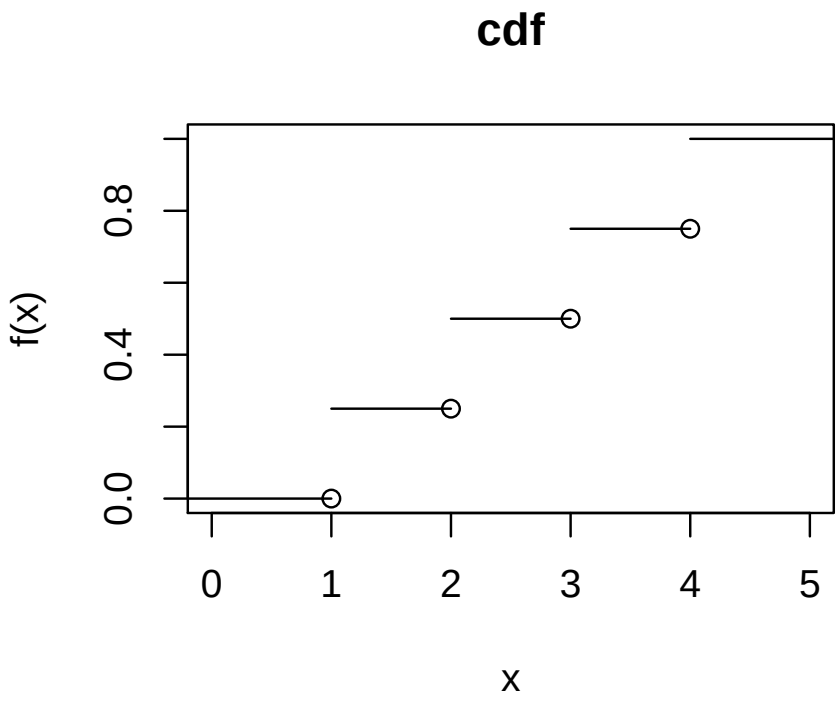
Example 9

Consider rolling a four-sided dice that produces numbers from 1 to 4 (so $X \sim \text{DUNIF}(1,4)$). Compute the cdf of X and plot it.

```
(U <- RV(1:4, probs = rep(1/4, 4)))
```

```
## Random variable with 4 outcomes
##
## Outcomes   1   2   3   4
## Probs    1/4 1/4 1/4 1/4
```

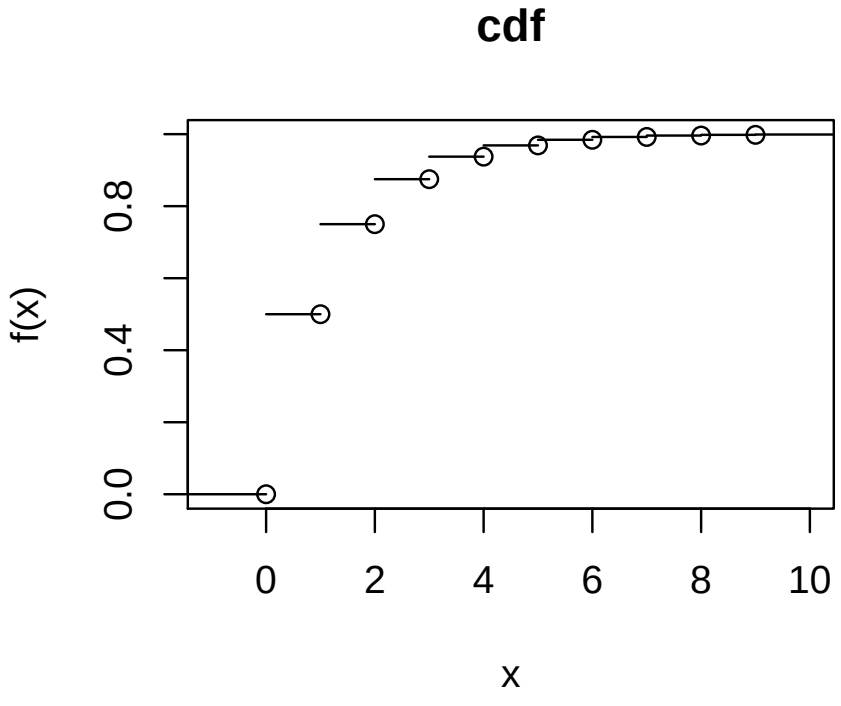
```
discunifcdf <- function(u) {P(U <= u)}
discunifcdf <- Vectorize(discunifcdf)
plot(stepfun(1:4, discunifcdf(0:4), right = TRUE), verticals = FALSE,
     main = "cdf")
```



Example 10

Find the cdf of a geometric random variable with parameter p. What would a plot of the cdf look like?

```r
geomcdf <- function(q) {pgeom(q, 1/2)}
plot(stepfun(0:9, geomcdf((-1):9), right = TRUE), vertical = FALSE,
     main = "cdf")
```



Section 3: Expected Values

The **expected value** for a discrete random variable, $\mathbb{E}[X]$, is given below:

$\mathbb{E}[X]$ is viewed as the population mean, μ, described in previous chapters. We can also compute the expected value of functions of X, $\mathbb{E}[h(X)]$, in a natural way:

The expected value is, in some sense, a "best prediction"[33] for the value of X.

[33] Specifically, let $\mathbb{E}\left[(X-\hat{\mu})^2\right]$ represent the expected squared error, and $\hat{\mu}$ represents a prediction for X; when this quantity exists (sometimes it doesn't), the value of $\hat{\mu}$ that minizes the expected squared error is $\hat{\mu} = \mathbb{E}[X]$.

Example 11

Compute the expected value for $X \sim \text{Ber}(p)$, $S \sim \text{DUNIF}(1, 6)$, and $N \sim \text{GEOM}(p)$ (as seen in examples 3, 4, and 5).

```
E(X)
```

```
## [1] 0.5
```

```
E(S)
```

```
## [1] 7
```

```
E(N)  # Approximate
```

```
## [1] 1.999999
```

Expectations are linear functions acting on random variables.[34]

[34] This is true for all random variables though we work with the discrete case only for now.

Proposition 8.

$$\mathbb{E}[aX + b] = a\mathbb{E}[X] + b$$

The **variance** of a random variable is given by:

There is a handy formula for computing the variance that is often easier than computing it directly:

Proposition 9.

$$\text{Var}(X) = \mathbb{E}\left[X^2\right] - (\mathbb{E}[X])^2$$

$\text{Var}(X)$ is thought of as the population variance and is often denoted $\text{Var}(X) = \sigma^2$. From this we get the population standard deviation, $\sigma = \sqrt{\sigma^2}$.

Example 12

Compute the variance and standard deviation of the random variables listed in Example 11.

```
V(X)
```

```
## [1] 0.25
```

```
SD(X)
```

```
## [1] 0.5
```

```
V(S)
```

```
## [1] 5.833333
```

```
SD(S)
```

```
## [1] 2.415229
```

```
V(N)   # Approximate
```

```
## [1] 1.999981
```

```
SD(N)
```

```
## [1] 1.414207
```

Proposition 10.

$$\operatorname{Var}(aX+b) = a^2 \operatorname{Var}(X)$$
$$\sigma_{aX+b} = |a|\,\sigma_X$$

where σ_X is the standard deviation of X.

Expectations need not be finite or even exist, as demonstrated in the following example:[35]

[35] This example is known as the St. Petersburg paradox, and is famous for how unintuitive its solution is and how strongly humans underestimate the game's value. The game gets its name due to its resolution by Daniel Bernoulli in 1738 (Bernoulli, 1954), who lived in St. Petersburg at the time.

Example 13

Consider a game where a fair coin is flipped until it lands heads-up. A player would earn \$1 if the game ends with 1 flip, \$2 if it ends with two flips, \$4 if it ends with three flips, \$8 if it ends with four flips, and so on. The "fair price" of a game corresponds to the game's expected payout. What, then, is the fair price to play this game?

Section 4: The Binomial Probability Distribution

A **binomial experiment** is an experiment that satisfies the following requirements:

1. The experiment consists of n Bernoulli trials that end in either in "success", S, or "failure", F.
2. The trials are independent.
3. For each trial, $\mathbb{P}(S) = 1 - \mathbb{P}(F) = p \in (0,1)$.

We can think of the outcome of an experiment as a sequence of S and F, such as $SSFSF$ (here, $n = 5$).

The **binomial random variable** associated with a binomial experiment counts the number of "successes" in the experiment: $X(\omega) = \{\#\text{ of } S \text{ in } \omega\}$; we write $X \sim \text{BIN}(n,p)$. For example, $X(SSFSF) = 3$.

We denote the pmf of X with $b(x;n,p)$. This is 0 for x that is not an integer from 0 to n. For $x \in \{0,1,\ldots,n\}$, it can be computed:

The cdf of X is given below:

$\mathbb{E}[X]$, $\text{Var}(X)$, and σ_X are given below:[36]

[36] We can compute these algebraically or with a probabilistic argument. The former is algebraically tedious while the latter is illuminating and easy. We revisit this in Chapter 5.

Select values of $B(x; n, p)$ are given in Table A.1 of the textbook.

Example 14

You flip a fair coin ten times.

1. What is the probability you see exactly 4 heads? (Do so without using a table.)

2. If $X \sim \text{BIN}(10, 0.5)$, compute $\mathbb{P}(4 < X \leq 6)$.

3. Compute $\mathbb{P}(2 \leq X \leq 4)$.

4. What is the probability you see more than 7 heads?

5. Compute $\mathbb{E}[X]$, $\mathrm{Var}(X)$, and σ_X.

```r
dbinom(4, size = 10, prob = 0.5)  # 1
```

```
## [1] 0.2050781
```

```r
pbinom(6, size = 10, prob = 0.5) - pbinom(4, size = 10, prob = 0.5)  # 2
```

```
## [1] 0.4511719
```

```r
pbinom(4, size = 10, prob = 0.5) - pbinom(2 - 1, size = 10, prob = 0.5)  # 3
```

```
## [1] 0.3662109
```

```r
1 - pbinom(7, size = 10, prob = 0.5)  # 4
```

```
## [1] 0.0546875
```

```r
pbinom(7, size = 10, prob = 0.5, lower.tail = FALSE)  # Alternative to 4
```

```
## [1] 0.0546875
```

```r
X <- RV(0:10, probs = dbinom(0:10, size = 10, prob = 0.5))
E(X)  # 5
```

```
## [1] 5
```

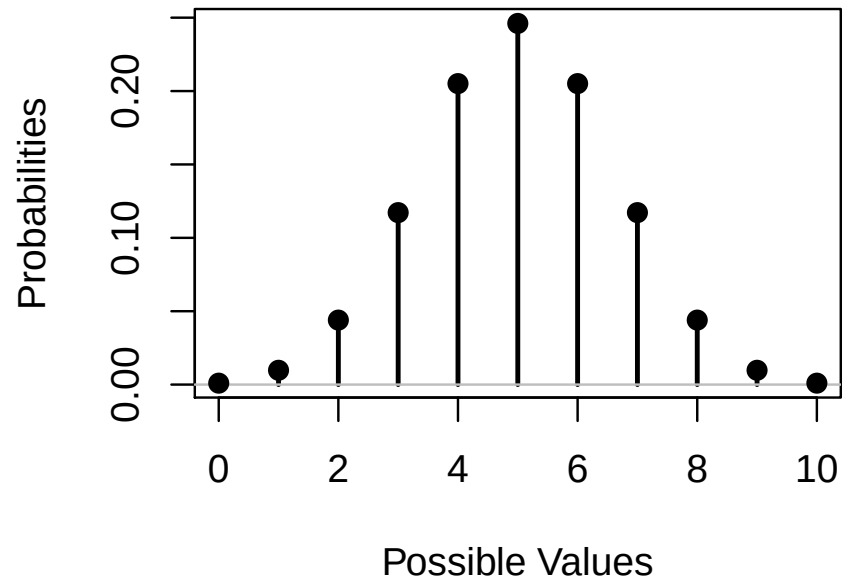
```r
V(X)
```

```
## [1] 2.5
```

```r
SD(X)
```

```
## [1] 1.581139
```

```r
plot(X)
```



Example 15

A manufacturer of widgets send batches of widgets in giant bins. Your company will accept a shipment of widgets if no more than 7% of widgets are defective. The procedure for deciding whether a shipment is defective is to choose four widgets from the batch at random, without replacement. If more than one widget is defective, the batch is rejected. What is the probability of rejecting the batch if 7% of the widgets are defective? Model the process using a binomial random variable.[37]

[37] We can view the batch of widgets as the entire population and we are choosing a subsample of that population without replacement. Binomial random variables draw "successes" and "failures" from an infinite population, not a finite one, and thus a different probability distribution should describe this experiment (it is the subject of the next section). It is safe, though, to treat a finite population like an infinite one if your sample size does not exceed 5% of the population size.

```r
pbinom(1, 4, 0.07, lower.tail = FALSE)
```

```
## [1] 0.02672803
```

Example 16

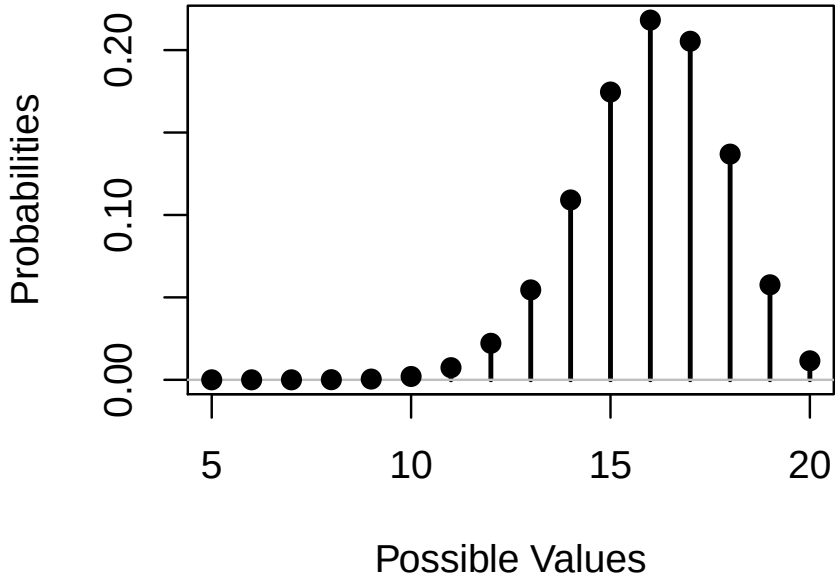
I claim that I can make 80% of my free-throw shots when playing basketball. You plan to test me by having me shoot 20 baskets; if I make fewer baskets than a specified amount, you will call me a liar. The threshold amount of baskets is chosen so that the probability I make less than this amount given that I am, in fact, an 80% free-throw shooter does not exceed 5%. What is the threshold amount?

Additionally, compute the mean and standard deviation of the number of shots I would make if my claim is true.

```
qbinom(0.05, 20, 0.8) - 1
```

```
## [1] 12
```

```
## The subtraction is due to how qbinom defines quantiles; see documentation
X <- RV(0:20, probs = dbinom(0:20, size = 20, prob = 0.8))
plot(X)
```




```
E(X)
```

```
## [1] 16
```

```
V(X)
```

```
## [1] 3.199998
```

```
SD(X)
```

```
## [1] 1.788854
```

Section 5: Hypergeometric and Negative Binomial Distributions

The **hypergeometric distribution** is the finite population analogue to the binomial distribution. The population has N elements labeled S or F (for "success" or "failure"). There are M S's in the population (and thus $N - M$ F's). A sample of size n is chosen from the sample without replacement in such a way that each subset of n elements is equally likely.[38] $X \sim \text{HYPERGEOM}(n, M, N)$ denotes a random

[38] The hypergeometric distribution should resemble the binomial distribution as N becomes large. In fact it can be shown that if M is replaced with M_N and $\frac{M_N}{N} \to p$ as $N \to \infty$, the hypergeometric distribution becomes the binomial distribution in the limit.

variable following the hypergeometric distribution.

For an integer x satisfying $\max(0, n - N + M) \leq x \leq \min(n, M)$, the pmf of the hypergeometric distribution is given below:

Below are $\mathbb{E}[X]$ and $\text{Var}(X)$:

Example 16

A manufacturer of widgets send batches of widgets in giant bins. Your company will accept a shipment of widgets if no more than 6% of widgets are defective. The procedure for deciding whether a shipment is defective is to choose four widgets from the batch at random, without replacement. If more than one widget is defective, the batch is rejected. The batch sent contains 50 widgets. What is the probability of rejecting the batch if 6% of the widgets are defective?

Also, compute the mean and variance of X, the number of defective widgets in the sample, under the assumption that 6% of widgets are defective.

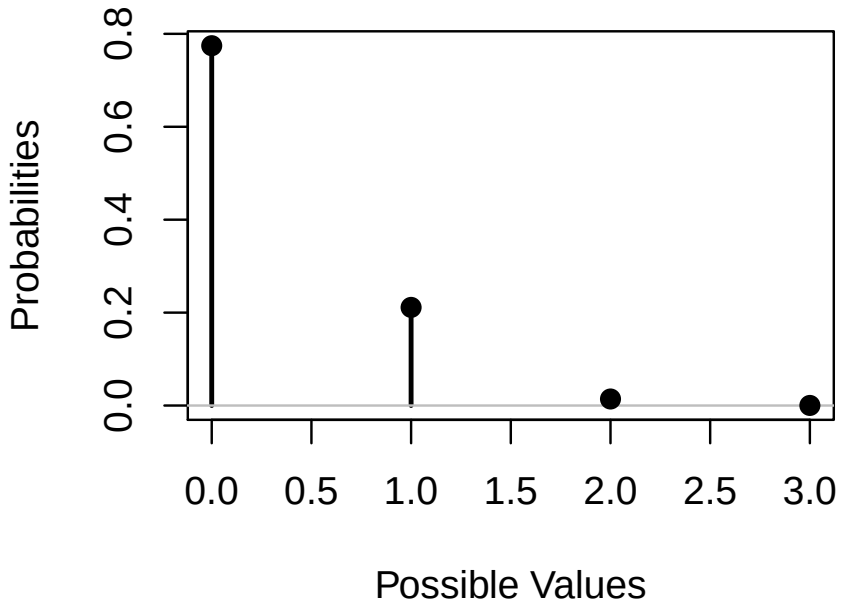
```
phyper(1, 50 * .06, 50 * (1 - .06), 4, lower.tail = FALSE)
```

```
## [1] 0.01428571
```

```
pbinom(1, 4, .06, lower.tail = FALSE)  # For comparison
```

```
## [1] 0.01991088
```

```
X <- RV(0:4, probs = dhyper(0:4, 50 * .06, 50 * (1 - .06), 4))
plot(X)
```




```
E(X)
```

```
## [1] 0.24
```

```
V(X)
```

```
## [1] 0.2117878
```

Example 17

It is election night in the small town of Studentsville, and Jack Johnson is running for mayor against bitter rival, John Jackson. Votes have been cast and are being counted. There are 1024 ballots cast and among the 200 ballots counted, 116 were cast for Jack Johnson. If the election were actually a tie, what would be the probability of observing 116 ballots or more cast for Jack Johnson? What does this say about who is likely winning the election?

```
phyper(116 - 1, 512, 512, 200, lower.tail = FALSE)
```

```
## [1] 0.007203238
```

Consider flipping a coin with probability p of landing heads-up (all flips independent). Flip a coin until r heads have been seen, and count the number of tails seen until the experiment ended; let the random variable X represent this count. Then X follows the **negative binomial distribution**, or $X \sim \text{NB}(r, p)$.[39] The pmf of X is given below:

[39] Let $r = 1$. Then $Y = 1 + X$ follows the geometric distribution, or $Y \sim \text{GEOM}(p)$.

Additionally, below are $\mathbb{E}[X]$ and $\text{Var}(X)$:

Example 18

A husband and wife plan to have children until they have exactly two boys; after this, they will stop attempting to have children. Assume that the probability of giving birth to a boy is 51%.

1. What is the probability they will have two girls before stopping attempting to have more children?

2. What is the probability they will need at least four children?

3. What is the expected number of children they will have? What is the variance of this random variable?

```r
dnbinom(2, 2, 0.51)  # 1
```

```
## [1] 0.18735
```

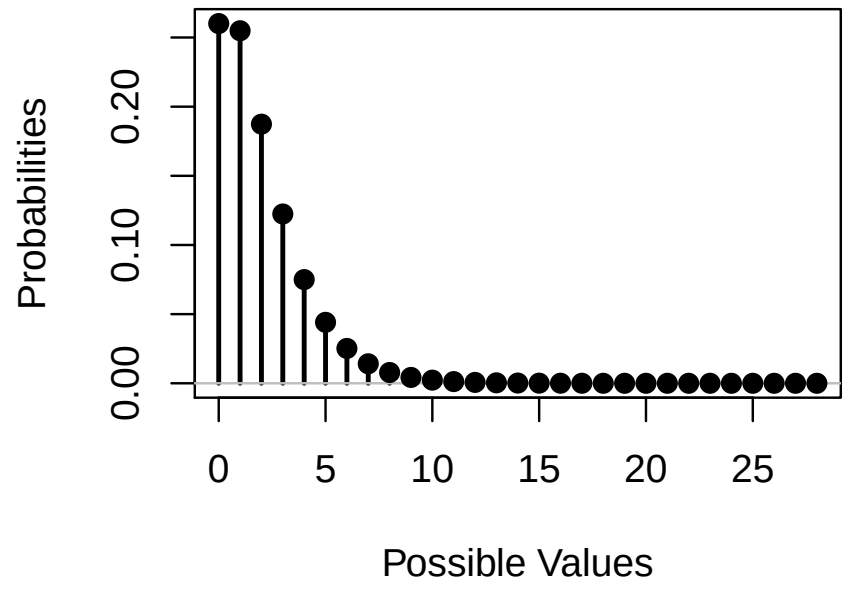
```r
pnbinom(4 - 2 - 1, 2, 0.51, lower.tail = FALSE)  # 2
```

```
## [1] 0.485002
```

```r
nbinom_func <- function(x) {dnbinom(x, 2, 0.51)}
(X <- RV(c(0, Inf), nbinom_func))
```

```
## Random variable with outcomes from 0 to Inf
##
## Outcomes     0     1     2     3     4     5     6     7     8     9    10    11
## Probs    0.260 0.255 0.187 0.122 0.075 0.044 0.025 0.014 0.008 0.004 0.002 0.001
##
## Displaying first 12 outcomes
```

```r
plot(X)
```




```r
E(X)  # 3
```

```
## [1] 1.921568
```

```r
V(X)
```

```
## [1] 3.767769
```

Section 6: The Poisson Probability Distribution

$X \sim \text{POI}(\mu)$, or X follows the Poisson distribution with parameter μ, if the pmf of X is given by:

Is this a valid pmf? Yes.

If $X \sim POI(\mu)$, $\mathbb{E}[X] = \text{Var}(X) = \mu$.

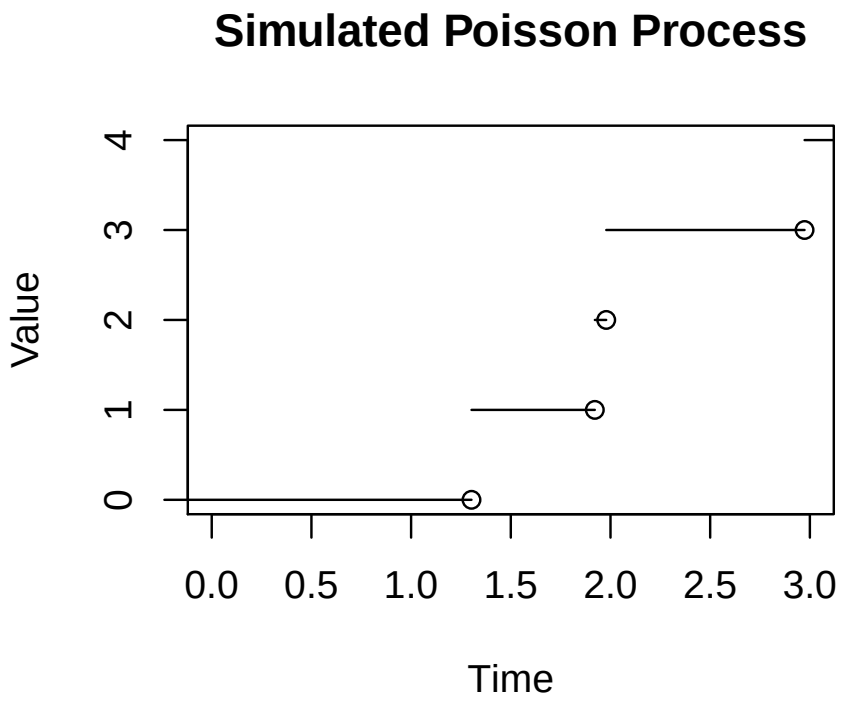
Table A.2 of your textbook gives the cdf of select Poisson distributions.

The Poisson distribution describes random variables that follow the Poisson process. This process describes the number of times an event occurs over an interval of time.[40] So the probability an event occurs k times during an interval of time of length t is given by $P_k(t) = \frac{e^{-\alpha t}(\alpha t)^k}{k!}$.

[40] This interpretation comes from a relationship between Poisson random variables and binomial random variables; if $p_n \to 0$ but $np_n \to \mu$ as $n \to \infty$, $b(x; n, p_n) \to p(x; \mu)$ as $n \to \infty$, where $p(\cdot; \mu)$ is the pmf of a Poisson random variable.

Below is a plot of a simulated Poisson process.

```
jumps <- rexp(4)
plot(stepfun(cumsum(jumps), 0:4, right = TRUE), vertical = FALSE,
     xlab = "Time", ylab = "Value", main = "Simulated Poisson Process",
     xlim = c(0, ceiling(sum(jumps))))
```



Example 19

On average, your daughter's soccer team scores 10 points per game. Assume that the number of points scored per game by her team follows a Poisson process.

1. What is the probability her team will score 7 points during the game?

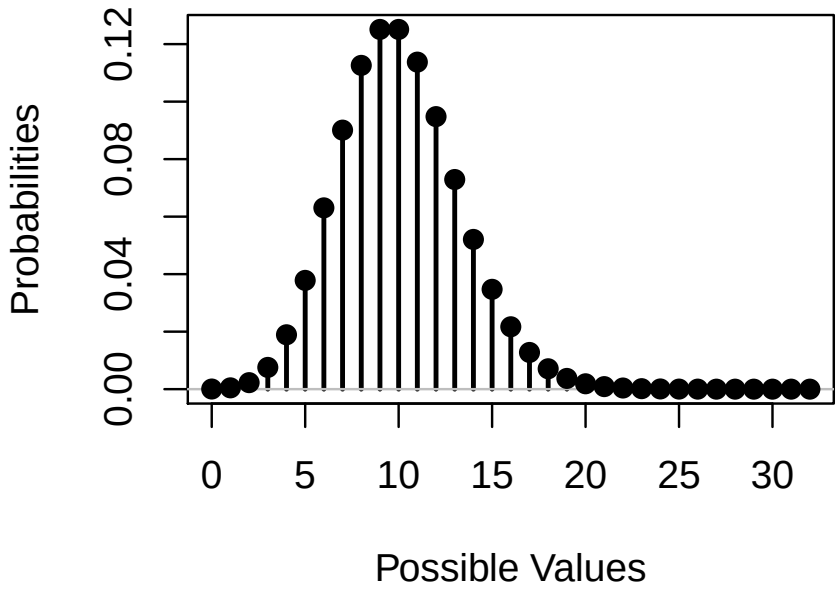
2. What is the probability that by half time your daughter's team will have scored two points?

3. What is the probability your daughter's team will score between 3 and 6 points (inclusive) during the game?

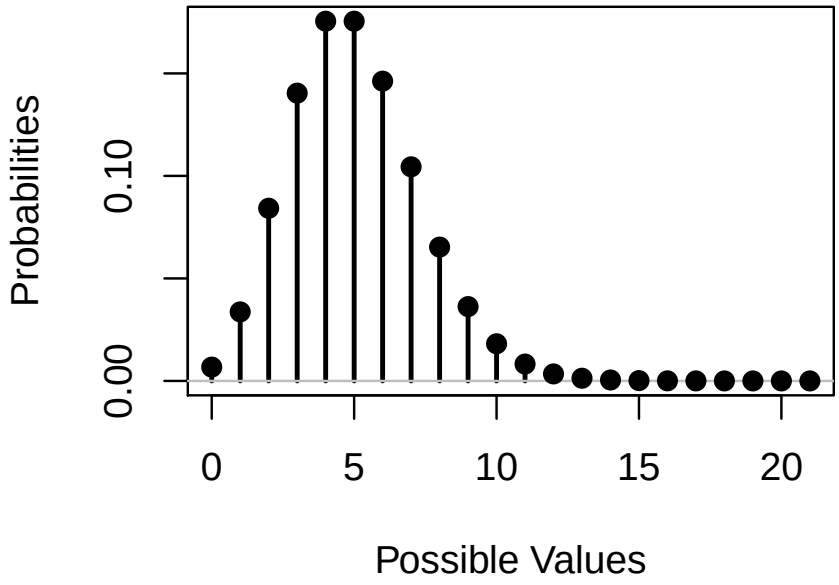
4. What is the probability your daughter's team will score more than three points in the last half of the game?

5. What is the probability that in two games your daughter's team will score between 15 and 18 points (inclusive)?

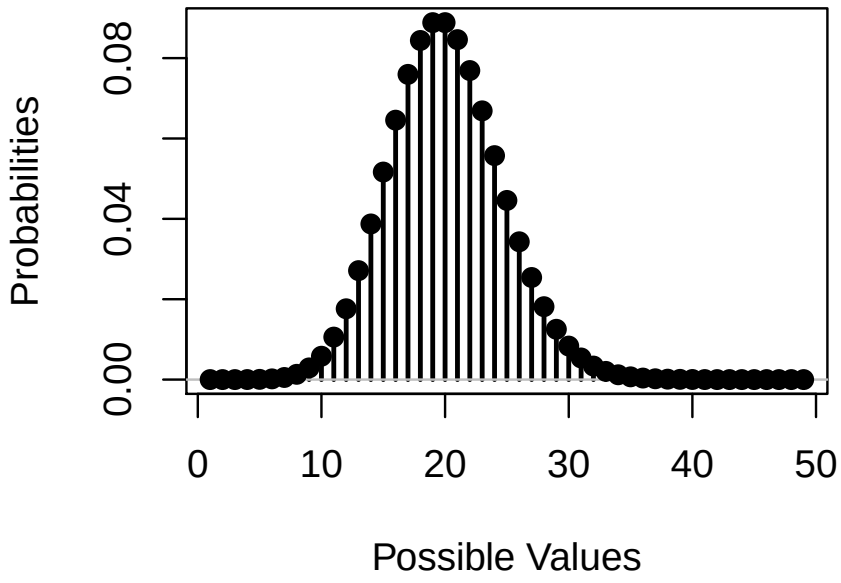
```
poiproc <- function(x, t) {dpois(x, 10 * t)}
(X1 <- RV(c(0, Inf), poiproc, t = 1))

## Random variable with outcomes from 0 to Inf
##
## Outcomes     0     1     2     3     4     5     6     7     8     9    10    11
## Probs    0.000 0.000 0.002 0.008 0.019 0.038 0.063 0.090 0.113 0.125 0.125 0.114
##
## Displaying first 12 outcomes

(Xhalf <- RV(c(0, Inf), poiproc, t = 1/2))

## Random variable with outcomes from 0 to Inf
##
## Outcomes     0     1     2     3     4     5     6     7     8     9    10    11
## Probs    0.007 0.034 0.084 0.140 0.175 0.175 0.146 0.104 0.065 0.036 0.018 0.008
##
## Displaying first 12 outcomes

(X2 <- RV(c(0, Inf), poiproc, t = 2))

## Random variable with outcomes from 0 to Inf
##
## Outcomes     1     2     3     4     5     6     7     8     9    10    11    12
## Probs    0.000 0.000 0.000 0.000 0.000 0.000 0.001 0.001 0.003 0.006 0.011 0.018
##
## Displaying first 12 outcomes

plot(X1)
```



```
plot(Xhalf)
```




```
plot(X2)
```




```
P(X1 == 7)  # 1
```

```
## [1] 0.09007923
```

```
P(Xhalf == 2)  # 2
```

```
## [1] 0.08422434
```

```
P((X1 >= 3) %AND% (X1 <= 6))  # 3
```

```
## [1] 0.127372
```

```
P(Xhalf > 3)  # 4
```

```
## [1] 0.7349741
```

```
P((X2 >= 15) %AND% (X2 <= 18))  # 5
```

```
## [1] 0.2765577
```

Tables for the Poisson distribution can be used for approximating binomial distribution probabilities when n is large and p is small. Then $b(x; n, p) \approx p(x; np)$.

Example 20

Use the Poisson approximation to estimate $B(4; 200, .01)$.

```
pbinom(4, 200, .01)
```

```
## [1] 0.9482537
```

```
ppois(4, 200 * .01)
```

```
## [1] 0.947347
```

Chapter 4: Continuous Random Variables and Probability Distributions

Introduction

Continuous probability models are the other major class of probability models. In addition to extending our probabilistic framework to continuous phenomena (namely, measurements), the Normal[41] distribution is both a continuous distribution and arguably the most important distribution in statistics and probability theory, due to its role in the central limit theorem. Many of the concepts we covered for discrete random variables carry over to the continuous case, including pmfs (although they become density functions rather than mass functions), cdfs, and expectations. In fact, the continuous case may be slightly easier than the discrete case since $\mathbb{P}(X = c) = 0$ for all $c \in \mathbb{R}$ and $\mathbb{P}(X < x) = \mathbb{P}(X \leq x)$.

[41] Another name for the Normal distribution is the Gaussian distribution, named after the great mathematician Carl Friedrich Gauss. No one is sure where the name "Normal" came from, but some theorize that the distribution attracted so much attention authors began to refer to it as the "typical" distribution, although most natural phenomena doesn't follow a Normal distribution. Thus I capitalize the word "Normal" to refer to a particular distribution but as a reminder that the distribution doesn't automatically describe a phenomenon.

Section 1: Probability Density Functions

The analogue to the probability mass function seen for discrete random variables is the **probability density function (pdf)**. The pdf is a non-negative function $f(x)$ such that, for any two numbers a and b with $a \leq b$

In order for f to be a valid pdf we must also have

Example 1

Confirm that the function

$$f(x;a,b) = \begin{cases} \frac{1}{b-a} & a \leq x \leq b \\ 0 & \text{otherwise} \end{cases}$$

is a valid pdf. Then, plot the pdf. A random variable U following this distribution is said to follow the uniform distribution, denoted by $U \sim \text{UNIF}(a,b)$.

Example 2

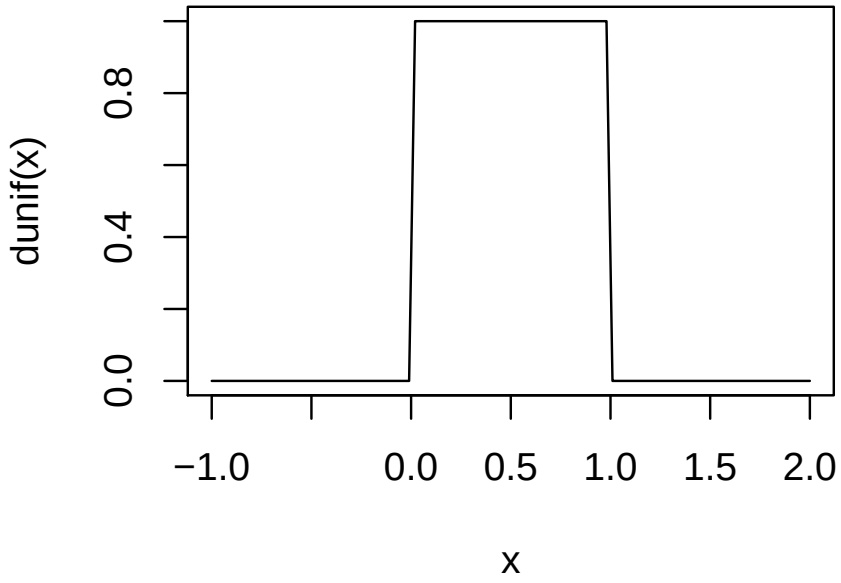
Confirm that the function

$$f(x;\mu) = \begin{cases} \frac{1}{\mu} e^{-\frac{1}{\mu}x} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

is a valid pdf. Then, plot the pdf. A random variable X following this distribution is said to follow the exponential distribution, denoted by $X \sim \text{EXP}(\mu)$[42].

[42] This notation is *not* standard and depends ultimately on who is writing the document. It turns out that μ is the mean of the exponential random variable when specified this way, but an alternative specification uses the rate $\lambda = \frac{1}{\mu}$. While the rate is often easier to work with mathematically, statisticians usually are interested in the mean. As a result, probabilists usually specify exponential random variables using the rate and write $X \sim \text{EXP}(\lambda)$ while statisticians prefer to specify exponential random variables using the mean. I do the latter as this is a statistics course, but be aware of the controversy.

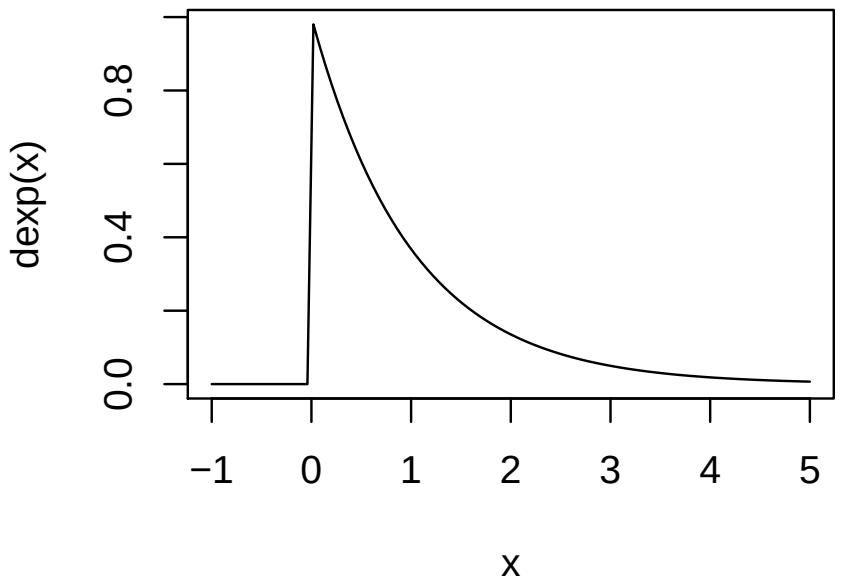
```r
## UNIF(0,1)
curve(dunif, -1, 2)      # Plot the pdf
```




```r
integrate(dunif, -1, 2)  # Integrage (numerically) the pdf to see it is one
```

```
## 1 with absolute error < 1.1e-15
```

```r
## EXP(1)
curve(dexp, -1, 5)
```




```r
integrate(dexp, 0, Inf)
```

```
## 1 with absolute error < 5.7e-05
```

Example 3

Accidents along a certain stretch of road are presumed to occur a distance of X miles from the nearest city center, where $X \sim \text{UNIF}(100, 150)$. Compute

1. $\mathbb{P}(110 \leq X \leq 130)$

2. $\mathbb{P}(127 < X \leq 144)$

3. $\mathbb{P}(X > 148)$

```r
integrate(dunif, 110, 130, min = 100, max = 150)  # 1
```

```
## 0.4 with absolute error < 4.4e-15
```

```r
integrate(dunif, 127, 144, min = 100, max = 150)  # 2
```

```
## 0.34 with absolute error < 3.8e-15
```

```r
integrate(dunif, 148, Inf, min = 100, max = 150)  # 3
```

```
## 0.03999993 with absolute error < 0.00011
```

Example 4

The time (in minutes) taken by a worker at the Tuition and Financial Aid office of a certain university to service a student follows an exponential distribution with $T \sim \text{Exp}(10)$. Compute the following:

1. $\mathbb{P}(T < 20)$

2. $\mathbb{P}(6 < T < 9)$

3. $\mathbb{P}(T \geq 22)$

```
integrate(dexp, -Inf, 20, rate = 1/10)  # 1
```

```
## 0.8646644 with absolute error < 3.8e-05
```

```
integrate(dexp, 6, 9, rate = 1/10)      # 2
```

```
## 0.142242 with absolute error < 1.6e-15
```

```
integrate(dexp, 22, Inf, rate = 1/10)   # 3
```

```
## 0.1108032 with absolute error < 1.3e-05
```

Section 2: Cumulative Distribution Functions and Expected Values

The cdf of a continuous random variable is

Thanks to the fundamental theorem of calculus we have the following relationship between the pdf and cdf of a random variable:

Rules for using the cdf to compute the probability of a continuous random variable taking values in an interval are given below.

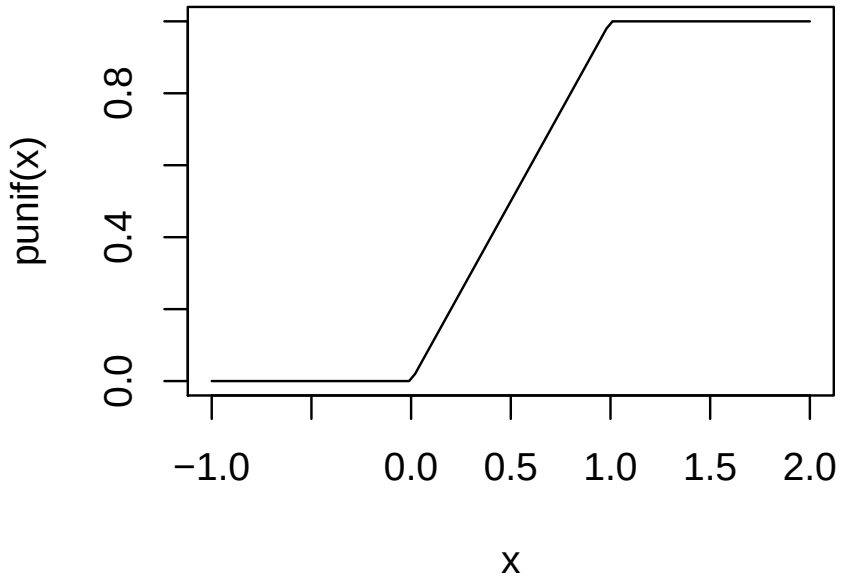
Example 5

Compute the cdf of $X \sim \text{UNIF}(a, b)$ and plot it.

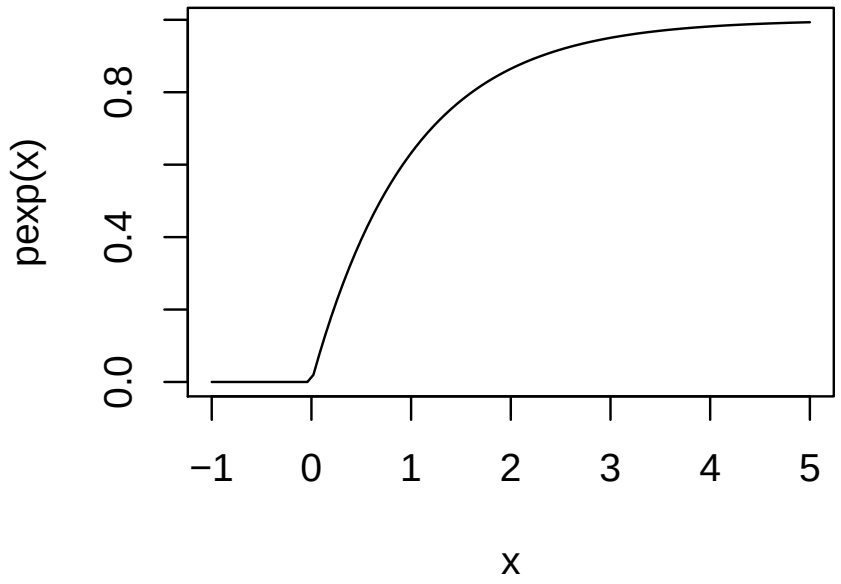
Example 6

Compute the cdf of $X \sim \text{EXP}(\mu)$ and plot it.

```r
curve(punif, -1, 2)  # CDF of UNIF(0, 1)
```




```r
curve(pexp, -1, 5)   # CDF of EXP(1)
```



Example 7

Answer the questions posed in Example 3 and Example 4 but using the cdf of the respective random variables.

```
## Example 3
punif(130, min = 100, max = 150) - punif(110, min = 100, max = 150)  # 1
```

```
## [1] 0.4
```

```
punif(144, min = 100, max = 150) - punif(127, min = 100, max = 150)  # 2
```

```
## [1] 0.34
```

```
1 - punif(148, min = 100, max = 150)                                 # 3
```

```
## [1] 0.04
```

```
## Example 4
pexp(20, rate = 1/10)                        # 1
```

```
## [1] 0.8646647
```

```
pexp(9, rate = 1/10) - pexp(6, rate = 1/10)  # 2
```

```
## [1] 0.142242
```

```
1 - pexp(22, rate = 1/10)                    # 3
```

```
## [1] 0.1108032
```

The $100p$**th percentile** (also referred to as quantiles) of a distribution is the number $\eta(p)$ such that $F(\eta(p)) = p$. If F can be inverted over its support, we can use F^{-1} to find percentiles.

A particularly interesting percentile is the 50th percentile, otherwise known as the **median**, $\tilde{\mu}$.

Example 8

Find percentile functions for the uniform and exponential distributions. Then find $\eta(0.5)$.

```
## Example for UNIF(0,1) and EXP(1)

qunif(0.5)

## [1] 0.5

qexp(0.5)

## [1] 0.6931472
```

Below are formulas for $\mathbb{E}[X]$, $\mathbb{E}[h(X)]$, and $\text{Var}(X)$ in the continuous case.

The shortcut formula for the variance in the discrete case also holds in the continuous case.

Proposition 11.

$$\text{Var}(X) = \mathbb{E}\left[X^2\right] - (\mathbb{E}[X])^2$$

Example 9

Compute $\mathbb{E}[X]$ and $\text{Var}(X)$ for uniform and exponential random variables.

```
(mu1 <- integrate(function(x) {x * dunif(x, 0, 1)}, -1, 2))  # Mean of UNIF(0,1)

## 0.5 with absolute error < 5.6e-16

integrate(function(x) {(x - mu1$value)^2 * dunif(x, 0, 1)},  # Var of UNIF(0,1)
          -1, 2)

## 0.08333333 with absolute error < 8.6e-05

(mu2 <- integrate(function(x) {x * dexp(x)}, 0, Inf))         # Mean of EXP(1)

## 1 with absolute error < 6.4e-06

integrate(function(x) {(x - mu2$value)^2 * dexp(x)}, 0, Inf)  # Var of EXP(1)

## 1 with absolute error < 5.8e-05
```

Section 3: The Normal Distribution

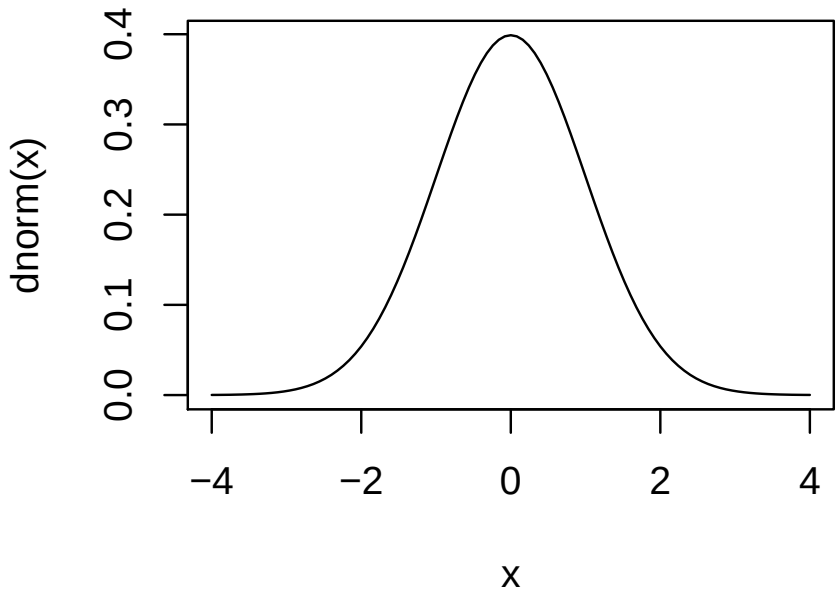
We say that a random variable X follows the **Normal distribution**[43], or $X \sim N(\mu, \sigma)$[44], if it has the pdf:

Below is a sketch of the density curve for the Normal distribution:

[43] Of all the probability distributions, the Normal distribution is arguably the most important. It plays a prominent role in one of the key theorems of probability, the central limit theorem, and as a result many random variables start to resemble Normally distributed random variables under certain conditions; we will see examples in this section. It is a well-behaved distribution; while any real number could be generated by the Normal distribution, it is effectively supported on the interval $[\mu - 3\sigma, \mu + 3\sigma]$. It naturally describes phenomena we would say results from an error process. That said, not everything is Normally distributed. Stock price movements, for example, are modeled with the Normal distribution yet we see fluctuations that would never be seen in billions of years if the Normal distribution were actually the appropriate distribution.

[44] Frequently the Normal distribution is specified with σ^2 instead of σ. In this class we use σ, but be aware that in academic settings it may be more common to see the Normal distribution using σ^2 instead. This is because the math is generally easier when using σ^2 and the notation extends well to multivariate or even functional cases.

```r
curve(dnorm, -4, 4)  # Plot of the density curve for N(0,1)
```



$\mathbb{E}[X]$, $\text{Var}(X)$, and $\text{SD}(X)$ are given below.

One property of the Normal distribution is the **68-95-99.7 rule**:

If $Z \sim N(0,1)$, we say that Z follows the **standard Normal distribution**. This distribution is useful since we can relate $X \sim N(\mu, \sigma)$ to the standard Normal distribution, and vice versa:

Let $\Phi(z) = \mathbb{P}(Z \leq z)$ be the cdf of the standard Normal distribution. Then if $F(x) = \mathbb{P}(X \leq x)$, we have the following relationship between F and Φ:

This means that we only need to worry about tabulating values for $\Phi(z)$[45] for working with any Normal distribution, as done in Table A.3.

Example 10

Compute the following:

1. $\mathbb{P}(Z \leq 0)$

2. $\mathbb{P}(Z \leq 1.23)$

[45] Notice that

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{t^2}{2}} dt$$

As mentioned, there is no closed form solution to this integral, but that is not a problem. Numerical methods can easily compute these quantities and they can then be tabulated. On a more general note, we encounter integrals without closed form solutions all the time, yet the functions they represent can still be very well-behaved, so there is no problem leaving the integral in the expression of the quantity; we know the integral exists, we can evaluate it numerically, and we can even talk about its properties. Not every integral needs to be like the ones seen in the calculus sequence of classes.

```
pnorm(0)     # 1
```

```
## [1] 0.5
```

```
pnorm(1.23)  # 2
```

```
## [1] 0.8906514
```

3. $\mathbb{P}\left(-1.97 \leq Z \leq 2.1\right)$

4. $\mathbb{P}\left(Z \geq 1.8\right)$

5. $\mathbb{P}\left(Z > 5.2\right)$

```r
pnorm(2.1) - pnorm(-1.97)       # 3
```

```
## [1] 0.9577164
```

```r
1 - pnorm(1.8)                  # 4
```

```
## [1] 0.03593032
```

```r
pnorm(5.2, lower.tail = FALSE)  # 5
```

```
## [1] 9.964426e-08
```

Example 11

IQ scores are said to be Normally distributed with mean 100 and standard deviation 15. Let Q be a randomly selected individual's IQ score. Compute the following:

1. $\mathbb{P}(85 \leq Q \leq 115)$

2. $\mathbb{P}(Q > 90)$

3. The International Society for Philosophical Enquiry requires potential members to have an IQ of at least 135 in order to join the society. Based on this, what proportion of the population is eligible for membership?

```r
pnorm(115, mean = 100, sd = 15) - pnorm(85, mean = 100, sd = 15)  # 1
```

```
## [1] 0.6826895
```

```r
pnorm(90, mean = 100, sd = 15, lower.tail = FALSE)                # 2
```

```
## [1] 0.7475075
```

```r
pnorm(135, mean = 100, sd = 15, lower.tail = FALSE)               # 3
```

```
## [1] 0.009815329
```

Here the notation z_α is used to mean $\Phi(z_\alpha) = 1 - \alpha$. We can relate this back to general $\eta(p)$, defined for an arbitrary Normally distributed random variable.

$z_{1-\alpha}$ can be found using Table A.3 using a reverse lookup.

Example 12

1. What is $z_{0.5}$?

2. What is $z_{0.05}$?

3. What are the first and third quartiles of the standard Normal distribution?

```
qnorm(0.5)                      # 1
```

```
## [1] 0
```

```
qnorm(0.05, lower.tail = FALSE)  # 2
```

```
## [1] 1.644854
```

```
qnorm(0.25)                     # 3a
```

```
## [1] -0.6744898
```

```
qnorm(0.75)                     # 3b
```

```
## [1] 0.6744898
```

Example 13

Using the description of the random variable Q from Example 11, answer the following questions.

1. Mensa International require individuals have an IQ score that would place them in the top 2% of the population. What is the minimum IQ score needed to be a member of Mensa?

2. The part of the population with the lowest 5% of IQ scores is considered to be intellectually disabled. What is the highest IQ score needed to be in this group?

```r
qnorm(0.02, mean = 100, sd = 15, lower.tail = FALSE)  # 1
```

```
## [1] 130.8062
```

```r
qnorm(0.05, mean = 100, sd = 15)                      # 2
```

```
## [1] 75.3272
```

Due to the symmetry of the Normal distribution we have the following useful identities for Φ:

As mentioned before, Φ can be used to approximate the cdf of other random variables. Below is a particular example for binomial random variables when n is large[46]:

[46] A rule of thumb is that if $np \geq 10$ and $n(1-p) \geq 10$, it is safe to use this approximation.

Example 14

A manufacture will reject a batch of widgets if, in a sample of 100 randomly selected widgets from the batch, 15 or more are defective.

If 12% of the widgets in the batch are defective, what is the probability of rejecting the batch? (Use the Normal approximation to answer this question.)

```
1 - pnorm((15 + 0.5 - (.12 * 100))/sqrt(.12 * .88 * 100))
```

```
## [1] 0.1407288
```

The approximation works for Poisson random variables too, when λ is large; choose $\mu = \lambda$ and $\sigma = \sqrt{\lambda}$ for the approximation.[47]

[47] Many of the distributions we see can be related to the Normal distribution in some way.

Example 15

Suppose $X \sim \text{POI}(100)$. Estimate $\mathbb{P}(X \leq 110)$.

```r
pnorm(110 + 0.5, mean = 100, sd = sqrt(100))
```

```
## [1] 0.8531409
```

```r
ppois(110, 100)  # For comparison
```

```
## [1] 0.8528627
```

Section 4: The Exponential and Gamma Distributions

We have investigated the properties of the exponential distribution already; below we recall what we have seen:

Exponential random variables can be used to model waiting times, particularly when a process is **memoryless**; that is, the time remaining until the process terminates is independent of how long the process has currently taken.

Proposition 12 (Memoryless property)**.** *Let $T \sim EXP(\mu)$. Then*

$$\mathbb{P}(T \geq t + t_0 | T \geq t_0) = \mathbb{P}(T \geq t)$$

Exponential random variablest play an important role in Poisson processes. The time between subsequent jumps of a Poisson process with parameter α follow an exponential distribution with mean $\mu = \frac{1}{\alpha}$.

Example 16

Your daughter's team score on average 10 points per game. You model the points scored by her team in a game with a Poisson pro-

cess, and $t = 1$ is a whole game.

1. Based on this, what is the expected time between points score by your daughter's team?

2. Suppose that by the start of the second half your daughter's team has scored 3 points. Given this, what is the expected time when your daughter's team score is 4 points?

```
(mu3 <- integrate(function(x) {x * dexp(x, rate = 10)}, 0, Inf))  # 1
```

```
## 0.1 with absolute error < 4.9e-05
```

```
0.5 + mu3$value
```

```
## [1] 0.6
```

The **gamma function**, $\Gamma(\alpha)$, is given below:

The gamma function has interesting properties, including

(Based on this we can say that the gamma function is the continuous analogue to $n!$.)

The **(lower) incomplete gamma function**, $\gamma(\alpha, x)$, is given below:

This yields the obvious asymptotic relationship between $\gamma(\alpha, x)$ and $\Gamma(\alpha)$:

The following are the pdf and cdf of the **gamma distribution** with parameters α and β (we write $X \sim \text{GAMMA}(\alpha, \beta)$ to say X follows such a distribution):

If $\beta = 1$ then we refer to GAMMA$(\alpha, 1)$ as the **standard gamma distribution**. Table A.4 gives values of the cdf of the standard gamma distribution for particular α and x.

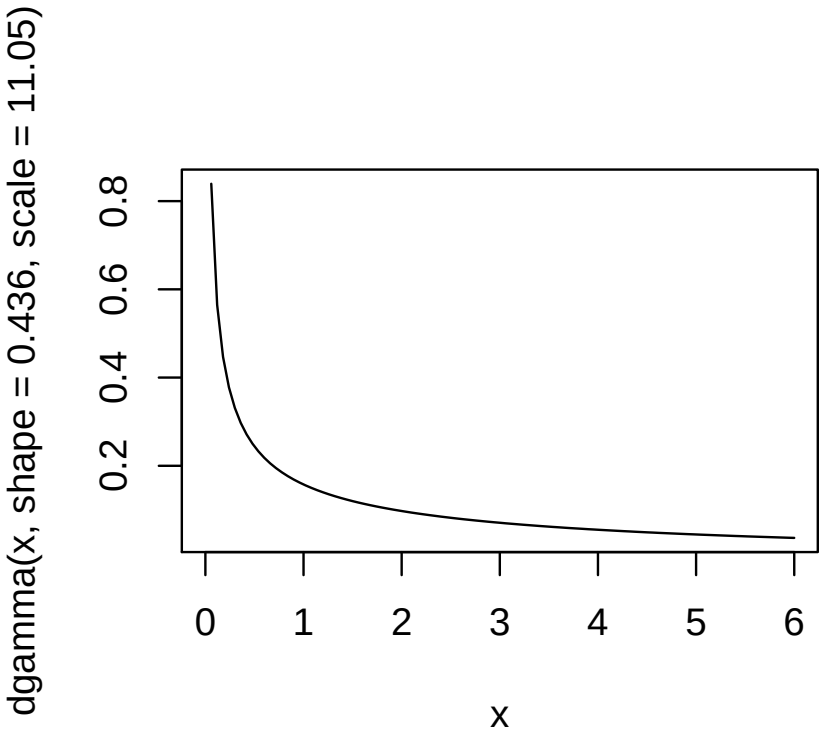
Standard gamma distributions can be used to compute probabilities involving non-standard gamma distributions in the following way:

The mean and variance of gamma-distributed random variables is given below:

Example 17

In a paper by Husak et al. (2007) the amount of rain (in mm) in Istanbull is fitted to a gamma distribution and the author estimated that the distribution of the amount of rain in April is $R \sim \text{GAMMA}(0.436, 11.05)$. Based on this, compute the mean and standard deviation of April rainfall.

```
curve(dgamma(x, shape = 0.436, scale = 11.05), 0, 6)
```




```
(mur <- integrate(function(x) {x * dgamma(x, shape = 0.436, scale = 11.05)},
                  0, Inf))
```

```
## 4.8178 with absolute error < 0.00029
```

```
(varr <- integrate(function(x) {(x - mur$value)^2 * dgamma(x,
                                                    shape = 0.436, scale = 11.05)},
                0, Inf))
```

```
## 53.23669 with absolute error < 0.0041
```

```
sqrt(varr$value)
```

```
## [1] 7.296348
```

```
## The probability the random variable is greater than 1
pgamma(1, shape = 0.436, scale = 11.05, lower.tail = FALSE)
```

```
## [1] 0.6145785
```

Let X_t be a Poisson process with rate parameter α. Let T_k be the time until the process is equal to k; that is, T_k is the smallest t such that $X_t = k$, so $X_{T_k} = k$. The distribution of T_k is known.

Example 18

You model the points your daughter's soccer team scores in a single game with a Poisson process with rate parameter $\alpha = 10$, with $t = 1$ representing a single game.

1. What is the mean and standard deviation of the time until her team scores 5 points?

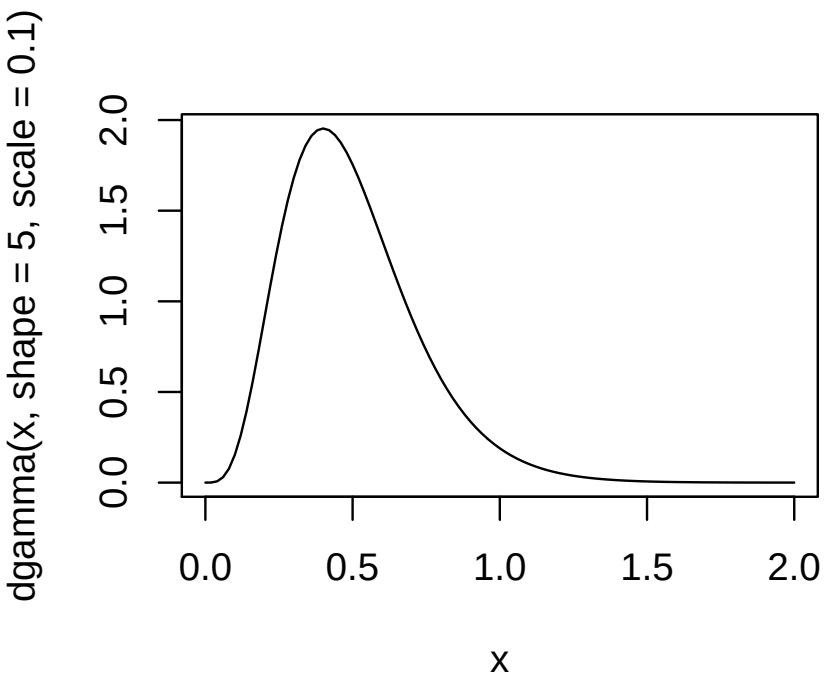
2. What is the probability that the time until her team scores 5 points is before half time ($t = 0.5$)?[48]

[48] Consider computing $\mathbb{P}(T_k > t)$, where t is a natural number. This is the event that the time when the process reaches k is greater than t; in other words, at time t, the process is less than k at time t, or $X_t < k$. Thus $\mathbb{P}(T_k > t) = \mathbb{P}(X_t < k)$. The left-hand side of this equality is an integral and the right-hand side is the sum $\sum_{x=0}^{k-1} \frac{e^{-\alpha t}(\alpha t)^x}{x!}$, so we have

$$\int_t^\infty \frac{\alpha^k x^{k-1} e^{-\alpha x}}{(k-1)!} dx = \sum_{x=0}^{k-1} \frac{e^{-\alpha t}(\alpha t)^x}{x!}$$

This identity could have been found with an inductive argument and integration by parts, but we have a probabilistic argument that explains why the identity holds, which is more illuminating.

```
curve(dgamma(x, shape = 5, scale = 0.1), 0, 2)
```




```
(mus <- integrate(function(x) {x * dgamma(x, shape = 5, scale = 0.1)},
                  0, Inf))
```

```
## 0.5 with absolute error < 3.5e-07
```

```
(vars <- integrate(function(x) {(x - mus$value)^2 * dgamma(x,
                                                    shape = 5, scale = 0.1)},
                0, Inf))
```

```
## 0.05 with absolute error < 2.7e-05
```

```
sqrt(vars$value)
```

```
## [1] 0.2236068
```

```
## The probability the random variable is greater than 1
pgamma(0.5, shape = 5, scale = 0.1)
```

```
## [1] 0.5595067
```

Notice that there is a relationship between the gamma distribution and the exponential distribution:

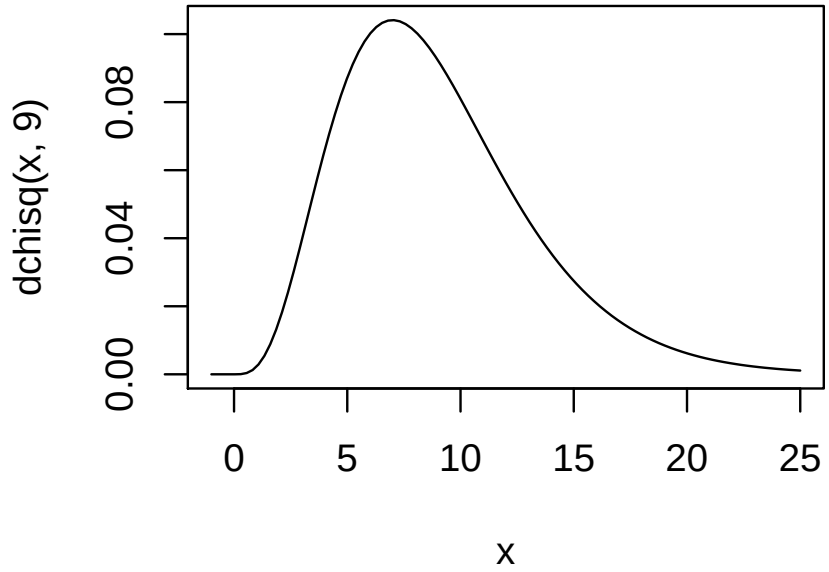
In this sense the exponential family of distributions is a subset of the gamma family of distributions.

The **chi-square distribution** is another distribution that belongs to the gamma family of distributions; we write $X \sim \chi^2(\nu)$ to indicate a chi-square distributed random variable. In particular, $X \sim \chi^2(\nu) \iff X \sim \text{GAMMA}(\nu/2, 2)$. This distribution is important in statistics for describing the sampling distribution of certain statistics. Values of the cdf of the chi-square distribution are given in Table A.7.

Example 19

Suppose $S^2 \sim \chi^2(9)$. Compute $\mathbb{E}\left[S^2\right]$, $\text{Var}\left(S^2\right)$, and $\mathbb{P}\left(S^2 > 3.325\right)$.

```r
curve(dchisq(x, 9), -1, 25)
```




```r
(mus2 <- integrate(function(x) {x * dchisq(x, 9)}, 0, Inf))
```

```
## 9 with absolute error < 7.6e-06
```

```r
integrate(function(x) {(x - mus2$value)^2 * dchisq(x, 9)}, 0, Inf)
```

```
## 18 with absolute error < 0.00012
```

```r
pchisq(3.325, 9, lower.tail = FALSE)
```

```
## [1] 0.9500055
```

Section 5: Other Continuous Distributions

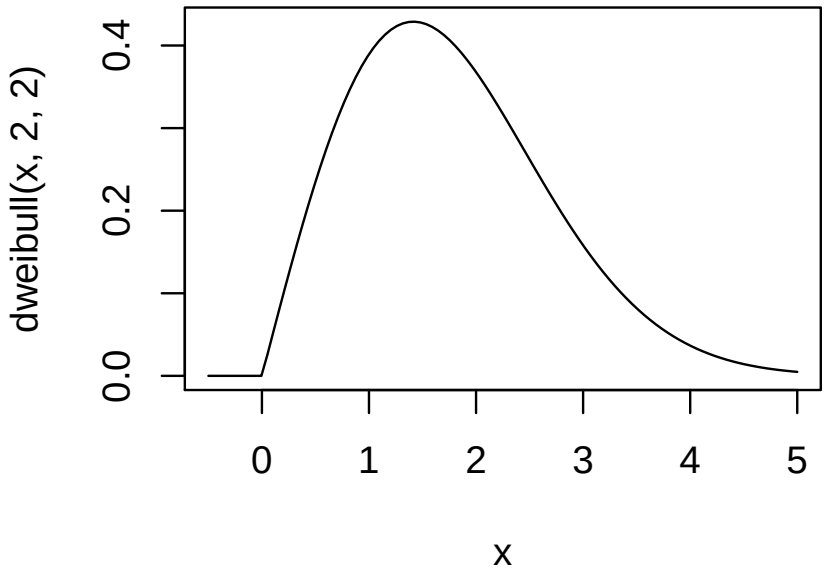
We say that X follows the **Weibull distribution** with shape parameter $\alpha > 0$ and scale parameter $\beta > 0$, or $X \sim \text{WEI}(\alpha, \beta)$[49], if the pdf of X is

[49] Sometimes $X \sim \text{WEI}(\alpha, \beta, \gamma)$ is seen, which means that $X - \gamma \sim \text{WEI}(\alpha, \beta)$; that is, X is a shifted version of the usual Weibull distribution.

The mean, variance, and cdf of the Weibull distribution are given below

If $\alpha = 1$ the Weibull distribution is an exponential distribution.
Below is a sketch of the pdf of the Weibull distribution

```r
curve(dweibull(x, 2, 2), -0.5, 5)
```



Example 20

Wind speed (in meters per second) at the site of a wind turbine is believed to follow a Weibull distribution with $\alpha = 2$ and $\beta = 8$. Compute the mean and median wind speeds and the standard deviation of wind speed.

The turbine will not turn if wind speed is below two meters per second. Compute the probability this occurs.

```
(muwind <- integrate(function(x) {x * dweibull(x, 2, 8)}, 0, Inf))
```

```
## 7.089815 with absolute error < 2.8e-06
```

```
(varwind <- integrate(function(x) {(x - muwind$value)^2 * dweibull(x, 2, 8)},
                      0, Inf))
```

```
## 13.73452 with absolute error < 6.6e-05
```

```
sqrt(varwind$value)
```

```
## [1] 3.706011
```

```
qweibull(0.5, 2, 8)  # Median
```

```
## [1] 6.660437
```

```
pweibull(2, 2, 8)
```

```
## [1] 0.06058694
```

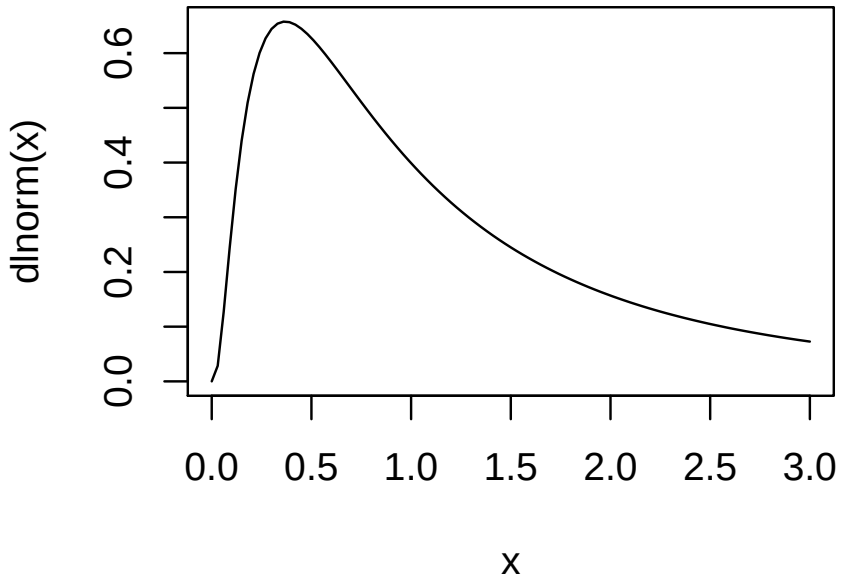
X is said to follow a **lognormal distribution**, denoted $X \sim LN(\mu, \sigma)$, if $\ln(X)$ follows a Normal distribution, or $\ln(X) \sim N(\mu, \sigma)$. X has pdf

We can express the cdf of X in terms of Φ like so:

μ and σ^2 are *not* the mean and variance of X. Instead we have

Below is a sketch of the pdf of X:

```r
curve(dlnorm, 0, 3)
```



Example 21

The current price of the stock with ticker symbol CGM is \$26.18. The quants believe the the price of the stock in a year is $Y = 26.18X$, where $X \sim LN(0.1, 0.2)$. Based on this information, find l and u such that $\mathbb{P}(l \leq Y \leq u) = 0.95$ and $\mathbb{P}(Y \leq l) = 0.025$.

```
(lprime <- qlnorm(0.025, 0.1, 0.2))
```

```
## [1] 0.7467739
```

```
(uprime <- qlnorm(0.975, 0.1, 0.2))
```

```
## [1] 1.635572
```

```
(l <- lprime * 26.18)  # lower bound
```

```
## [1] 19.55054
```

```
(u <- uprime * 26.18)  # upper bound
```

```
## [1] 42.81928
```

X follows the **beta distribution**, denoted $X \sim \text{BETA}(\alpha, \beta, A, B)$[50], if X has the pdf

[50] It is also common to see $X \sim \text{BETA}(\alpha, \beta)$, which refers to the standard beta distribution.

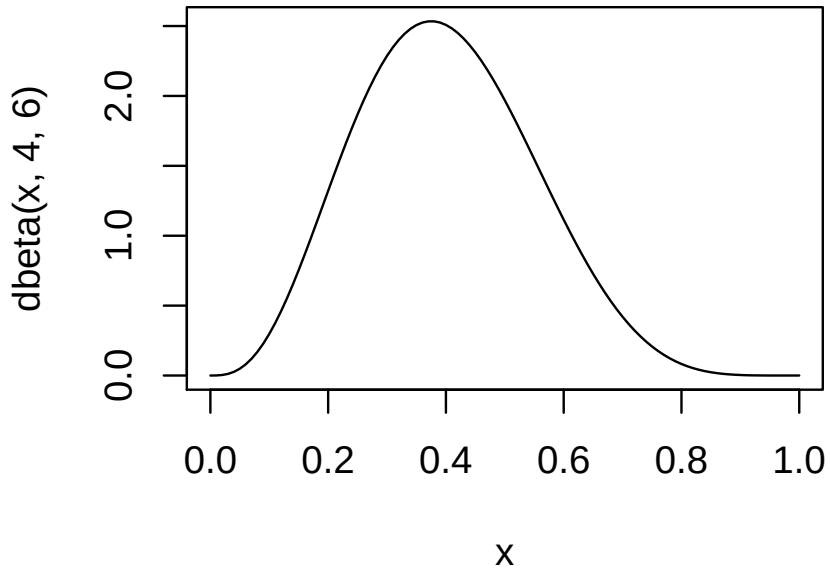
If $A = 0$ and $B = 1$, then X is said to have the **standard beta distribution**.

The mean and variance of X are given below:

The beta distribution can assume a large number of shapes depending on its shape parameters. But it has compact support, assigning positive probabilities only to regions between A and B.

Below is a sketch of what a beta distribution can look like.

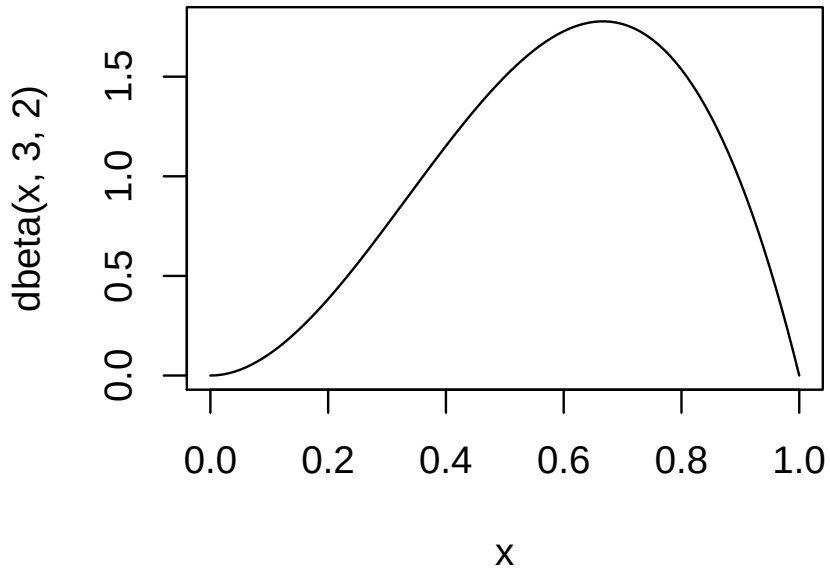
```r
curve(dbeta(x, 4, 6))
```



Example 22

Suppose $X \sim \text{BETA}(3,2)$. Write down the pdf of X and compute $\mathbb{E}[X]$, $\text{Var}(X)$, and $\mathbb{P}(1/4 \leq X \leq 3/4)$.

```
curve(dbeta(x, 3, 2))
```




```
(mux <- integrate(function (x) {x * dbeta(x, 3, 2)}, 0, 1))
```

```
## 0.6 with absolute error < 6.7e-15
```

```
(varx <- integrate(function (x) {(x - mux$value)^2 * dbeta(x, 3, 2)}, 0, 1))
```

```
## 0.04 with absolute error < 4.4e-16
```

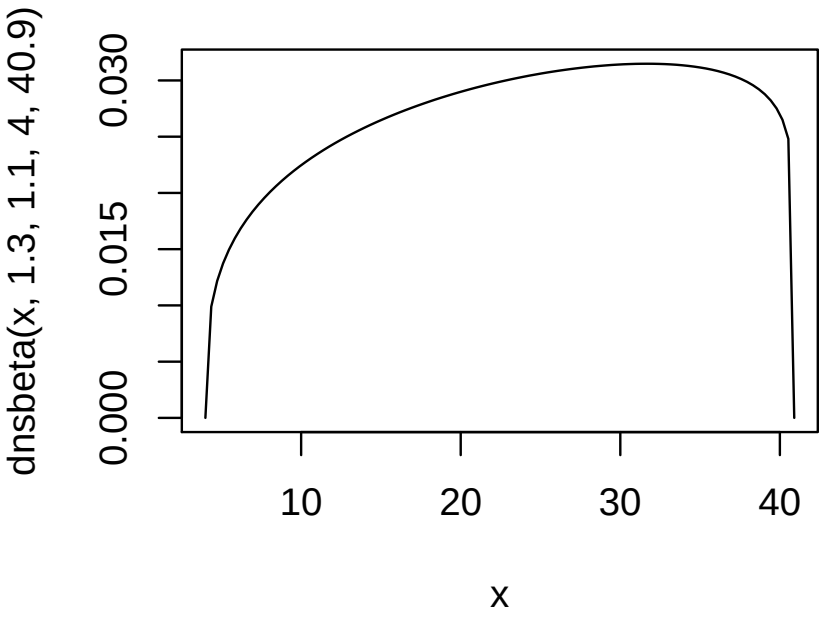
```
pbeta(.75, 3 ,2) - pbeta(.25, 3, 2)
```

```
## [1] 0.6875
```

Example 23

In a paper by Maltamo et al. (2007), the basal diameter (in cm) of pine trees was fitted to a beta distribution. The paper suggests that, if B is the diameter of a pine tree, then $B \sim \text{BETA}(1.3, 1.1, 4.0, 40.9)$. What, then, is the mean diameter of the pine trees? What about the standard deviation?

```r
suppressPackageStartupMessages(library(extraDistr))  # Package with more dist's

curve(dnsbeta(x, 1.3, 1.1, 4.0, 40.9), 4.0, 40.9)
```




```r
(mudiam <- integrate(function(x) {x * dnsbeta(x, 1.3, 1.1, 4.0, 40.9)},
                     4.0, 40.9))

## 23.9875 with absolute error < 1.4e-06

(vardiam <- integrate(function(x) {(x - mudiam$value)^2 * dnsbeta(x, 1.3, 1.1,
                                                                  4.0, 40.9)},
                      4.0, 40.9))

## 99.42312 with absolute error < 0.00012

sqrt(vardiam$value)

## [1] 9.971114
```

Section 6: Probability Plots

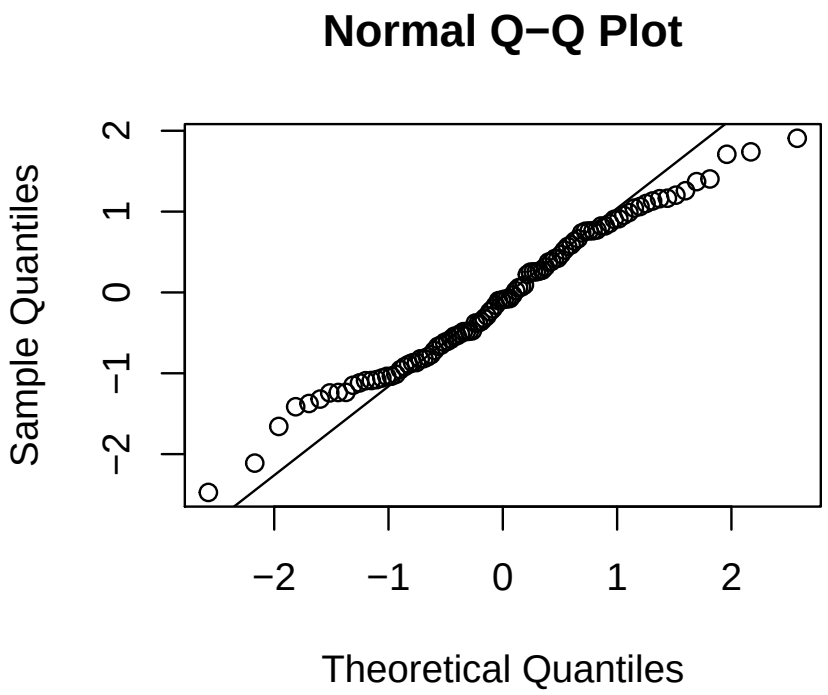
Probability plots are a visual method used to check whether a dataset could plausibly have been drawn from a particular distribution. In essence, we compare the observed sample percentiles with the percentiles of a dataset if it had come from a chosen distribution. If the relationship between the observed and the theoretical distributions is linear, the distributional assumption seems reasonable. If

there is a nonlinear relationship, the distribution chosen is not likely a good model for the data.

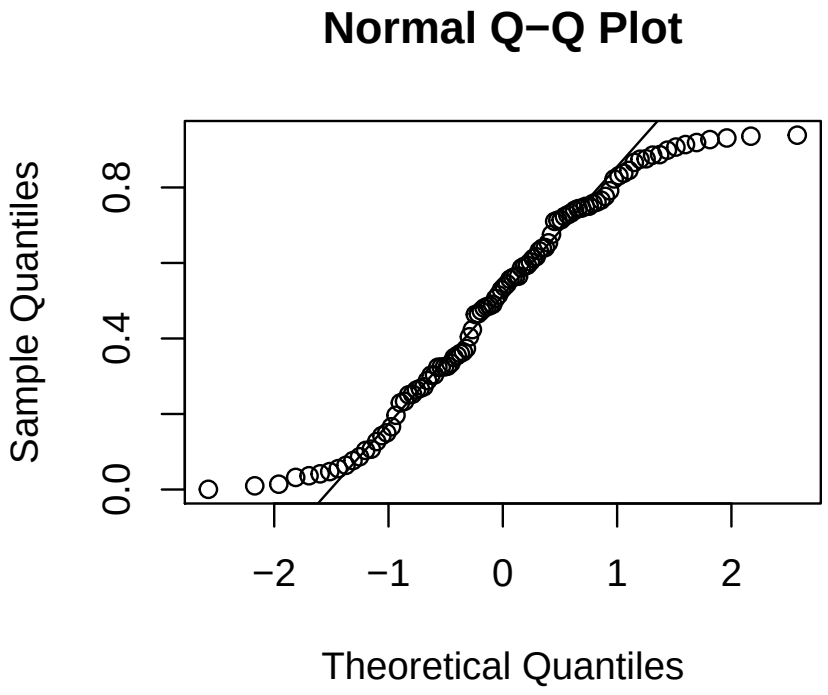
While we can often argue that a certain data generating process produces a particular probability distribution in the discrete case, fitting data to distributions is more difficult in the continuous case; we can't make arguments like we could in the discrete case. Thus we turn to probability plots or statistical tests.

Below is an example of a probability plot.

```r
dat1 <- rnorm(100)
dat2 <- runif(100)

## Probability plot checking for Normal distributions
qqnorm(dat1); qqline(dat1)
```




```r
qqnorm(dat2); qqline(dat2)
```



Suppose we have a sample $x_1, \ldots, x_n$, and $r_1, \ldots, r_n$ is the ordered sample (with $r_1 \leq r_2 \leq \ldots \leq r_n$). We call r_i the $[100(i-.5)/n]$**th sample percentile**.

To construct a probability plot, we do the following:

1. For $i = 1, \ldots, n$, we find the $[100(i-.5)/n]$th percentile of the *theoretical* distribution; we call these the theoretical percentiles, referred to as $\eta\left(\frac{i-0.5}{n}\right)$.
2. For $i = 1, \ldots, n$, plot the point $\left(\eta\left(\frac{i-0.5}{n}\right), r_i\right)$ on a Cartesian grid; the x-axis is the theoretical percentiles and the y-axis is the observed percentiles.

If the theoretical distribution is a Normal distribution, we call the probability plot a **Normal probability plot**.

We then decide if the relationship between the theoretical and observed percentiles appears linear. If yes, then the distribution is a good fit. Otherwise, it's a bad fit.

Example 24

Consider the following dataset:

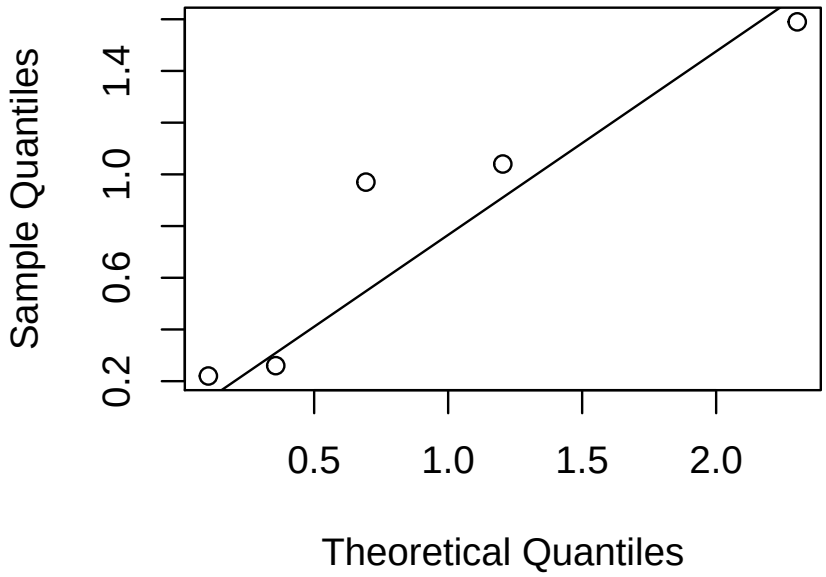
i	1	2	3	4	5
r_i	0.22	0.26	0.97	1.04	1.59

Create a probability plot to determine if it's plausible the data came from a EXP(1) distribution.

```r
x <- c(0.22, 0.26, 0.97, 1.04, 1.59)
(theo <- qexp((1:5 - 0.5)/5))
```

```
## [1] 0.1053605 0.3566749 0.6931472 1.2039728
## [5] 2.3025851
```

```r
qqplot(theo, x, xlab = "Theoretical Quantiles", ylab = "Sample Quantiles")
qqline(x, distribution = qexp)
```



As described we can check whether a dataset was generated by a *particular* distribution (in the last example, it was $\text{EXP}(1)$), but we usually want to know whether a dataset was generated by *a member of a family* of distributions (for example, $\text{EXP}(\mu)$). Fortunately there are tricks we can use to do the latter task.

We call θ_1 a **location** parameter and θ_2 a **scale** parameter if the cdf $F(x;\theta_1,\theta_2)$ depends on $\frac{x-\theta_1}{\theta_2}$[51]. Below are examples of parameters that are either (or are *neither*) location or scale parameters.[52]

[51] Intuitively, θ_1 shifts the pdf left or right rigidly, while θ_2 stretches or compresses the pdf.

[52] Notice that the mean is *not* always a location parameter. For the exponential distribution, the mean is a scale parameter.

If the theoretical distribution involes location and scale parameters, we estimate them; call the estimates $\hat{\theta}_1$ and $\hat{\theta}_2$. Instead of plotting using r_i, we use $\frac{r_i - \hat{\theta}_1}{\hat{\theta}_2}$, and use the *standard* theoretical distribution where $\theta_1 = 0$ and $\theta_2 = 1$[53].

Example 25

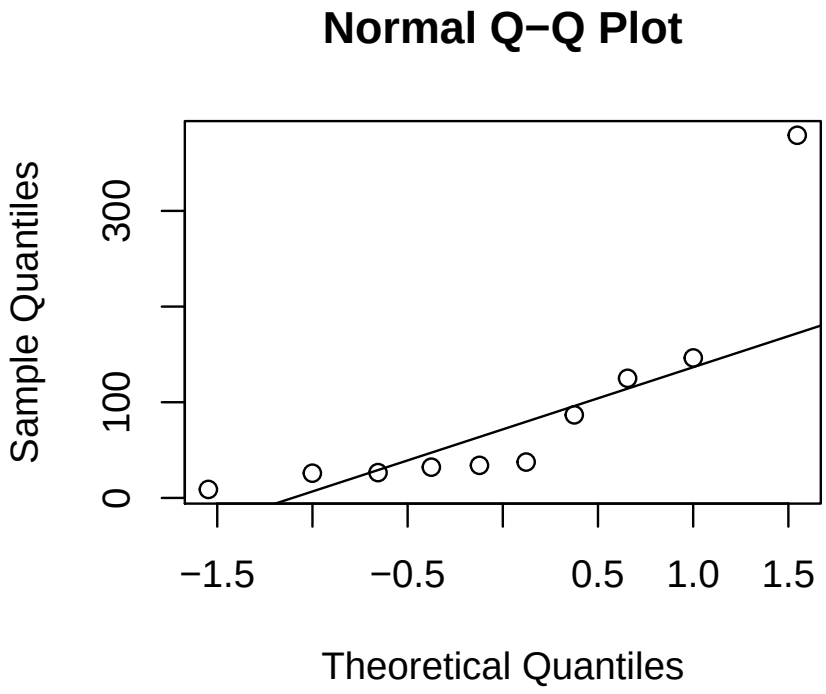
Construct a probability plot to check if the following dataset was plausibly generated by a Normal distribution.

i	1	2	3	4	5	6	7
r_i	8.89	25.86	26.47	32.16	34.07	37.49	86.80

i	8	9	10
r_i	125.02	146.36	379.06

[53] What if we need a parameter that is neither a location nor scale parameter? One trick would be to transform the data in an appropriate way. For example, if we think $X_i \sim LN(\mu, \sigma)$, neither μ nor σ are location or scale parameters, but we can create a probability plot for $\ln(X_i)$ instead and see if the new, transformed dataset is Normally distributed, as it should be if our hypothesis is correct; in this case, μ and σ can now be treated as location and scale parameters, respectively. This trick would not work if we wanted to check if $X_i \sim \text{BETA}(\alpha, \beta)$ since no transformation will turn α and β into location/scale parameters. In that case we may be forced to estimate α and β from the data, assuming that our hypothesis is true; in this example, call the estimates $\hat{\alpha}$ and $\hat{\beta}$. Then we would construct a probability plot to see if the data came from the distribution $\text{BETA}(\hat{\alpha}, \hat{\beta})$.

```r
y <- c(8.89, 25.86, 26.47, 32.16, 34.07, 37.49, 86.80, 125.02, 146.36, 379.06)
qqnorm(y)
qqline(y)
```



Chapter 5: Joint Probability Distributions and Random Samples

Introduction

We may naturally inquire about collections of random variables that are related to each other in some way. For instance, we may record an individual's height and weight, calling these random variables X and Y, and ask if these are correlated, uncorrelated, or even independent characteristics, and describe a probability model that accounts for the relationship in these two characteristics.

Additionally, we may have a large collection of random variables, say $X_1, X_2, \ldots, X_n$ which will be used to estimate some essential quantity of a distribution, such as the mean μ. We compute some quantity based on this collection of random variables, such as $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, or any other $T = T(X_1, \ldots, X_n)$. This quantity, dependent on random variables, is itself a random variable, and we call it a statistic. Being a statistic it has its own probability distribution, with its own mean and variance and cdf, and we can use the distribution of the statistic to make statements about the process that generated the original dataset $X_1, \ldots, X_n$. It is here when probability theory begins to turn into statistical theory.

Section 1: Jointly Distributed Random Variables

Suppose X and Y are two discrete random variables. Their **joint probability mass function** is described below:

This can be used to compute $\mathbb{P}\left((X, Y) \in A\right)$ for an event A:

From this we can compute the **marginal probability mass func-**

tions, $p_X(x)$ and $p_Y(y)$, for X and Y respectively.

These represent the probability distribution of X and Y respectively regardless of what value the other rv takes.

We can also compute what is known as the **conditional probability mass function of Y given $X = x$**, which represents the probability distribution of Y when we know that $X = x$. The **conditional probability mass function of X given $Y = y$** is defined in a similar manner.

Example 1

A fair six-sided die is rolled; let X represent the number of pips shown. At the same time, a fair coin is flipped, and $Y(\omega) = 1$ if the coin lands heads-up, and $Y(\omega) = 2$ if the coin lands tails-up. The joint pmf of X and Y is

1. Compute $\mathbb{P}(X < Y)$.

2. Compute $\mathbb{P}$ (Both X and $Y$are even)

3. Find the marginal pmfs for both X and Y.

4. Find the conditional distributions $p_{X|Y}(x|y)$ and $p_{Y|X}(y|x)$.

```r
library(discreteRV)
```

```
##
## Attaching package: 'discreteRV'
```

```
## The following object is masked from 'package:base':
##
##     %in%
```

```r
library(magrittr)  # Adds the %>% operator

XY <- jointRV(list(1:6, 1:2), probs = rep(1/12, times = 12))
(X <- marginal(XY, 1))  # The relationship between X and Y is still preserved
```

```
## Random variable with 6 outcomes
##
## Outcomes   1   2   3   4   5   6
## Probs    1/6 1/6 1/6 1/6 1/6 1/6
```

```r
(Y <- marginal(XY, 2))
```

```
## Random variable with 2 outcomes
##
## Outcomes   1   2
## Probs    1/2 1/2
```

```r
joint(X, Y)
```

```
## Random variable with 12 outcomes
##
## Outcomes  1,1  1,2  2,1  2,2  3,1  3,2  4,1  4,2  5,1  5,2  6,1  6,2
## Probs    1/12 1/12 1/12 1/12 1/12 1/12 1/12 1/12 1/12 1/12 1/12 1/12
```

```r
P(X < Y)
```

```
## [1] 0.08333333
```

```r
P((X %in% c(2, 4, 6)) %AND% (Y %in% c(2)))  # Both even
```

```
## [1] 0.25
```

```r
X | Y == 2  # Gets a conditional random variable
```

```
## Random variable with 6 outcomes
##
## Outcomes   1   2   3   4   5   6
## Probs    1/6 1/6 1/6 1/6 1/6 1/6
```

```
YgivenX <- function(x) {Y | X == x}
XgivenY <- function(y) {X | Y == y}

YgivenX(2)

## Random variable with 2 outcomes
##
## Outcomes   1   2
## Probs    1/2 1/2

XgivenY(2)

## Random variable with 6 outcomes
##
## Outcomes   1   2   3   4   5   6
## Probs    1/6 1/6 1/6 1/6 1/6 1/6
```

Now suppose that X and Y are continuous random variables. Much is the same; we work with a **joint probability density function**, **marginal probability density functions**, and **conditional probability density functions**.

Example 2

A company sells bags of "deluxe" mixed nuts, containing almonds, cashews, and peanuts. One bag is five pounds, and the joint pdf for the amount of almonds X and cashews Y in the bag (in pounds) is given below:

(We don't need to worry about the amount of peanuts; this is simply $5 - X - Y$ and thus is completely determined given X and Y.)

The region on which the pdf is illustrated below:

1. Customers buying bags of "deluxe" mixed nuts complain when 60% of the nuts in the bag are peanuts. Compute the probability this occurs.

2. Find the marginal distributions of X and Y. Use this to compute $\mathbb{E}[X]$, $\mathbb{E}[Y]$, $\text{Var}(X)$, and $\text{Var}(Y)$.

3. Find the conditional pdfs $f_{X|Y}(x|y)$ and $f_{Y|X}(y|x)$. Use this to compute $\mathbb{P}(X > 2|Y = 2)$.

We say that two random variables X and Y are **independent** if

---

Example 3

Are the random variables in the previous two example independent? Explain why or why not.

```r
independent(X, Y)  # For Example 1
```

```
## [1] TRUE
```

We can generalize our definitions from two to n rvs, $X_1, \ldots, X_n$.

We say that $X_1, \ldots, X_n$ are independent if for *every* subset $X_{i_1}, \ldots, X_{i_k}$ of the collection, we have

If $X_1, \ldots, X_n$ are independent and each has the same pmf/pdf that $X_1, \ldots, X_n$ are **independent and identically distributed**, often abbreviated *i.i.d.*.[54]

[54] This is a typical assumption about a dataset in statistics.

Example 4

Compute $\mathbb{P}\left(\min\{X_1, ..., X_n\} \geq x\right)$ if $X_1, \ldots, X_n$ are *i.i.d.*.

We can generalize the binomial distribution we saw before the the **multinomial distribution**. We have r categories, and a single observation belongs to category i with probability p_i. We count how many observations belong to category i; this gives X_i. Then the vector $(X_1, \ldots, X_r)$ follows the multinomial distribution, or $(X_1, \ldots, X_r) \sim \text{MULTINOM}(p_1, \ldots, p_r)$[55], and we have the pmf

[55] The binomial distribution is a particular instance of the multinomial distribution, when $r = 2$. We omit the count of tails, which we may call X_2, as it's redundant information given X_1.

Section 2: Expected Values, Covariance, and Correlation

Expectations involving two random variables are defined similarly to the univariate cases.

Example 5

Reconsider the random variables in Examples 1 and 2. Compute $\mathbb{E}[XY]$ for both cases.

```
E(X * Y)  # For Example 1's random variables
```

```
## [1] 5.25
```

One measure of the relationship between two random variables is the **covariance**.

The covariance is positive if the two random variables tend to be large together, while the covariance is negative if one rv tends to be large when the other tends to be small. If $\text{Cov}(X, Y) = 0$, then X and Y are **uncorrelated**.[56]

We also have the following shortcut formula for the covariance:

[56] Due to the relationship between the covariance and the variance, we sometimes see the notation $\text{Cov}(X, Y) = \sigma_{XY}$.

It's obtained in a similar manner to shortcut formulas found for computing $\text{Var}(X)$.[57]

Notice that the covariance is *not* insensitive to the units of the random variable; in fact, we can compute the covariance $\text{Cov}(aX + b, cY + d)$:

[57] From this it's clear that $\text{Cov}(X, X) = \text{Var}(X)$.

Changing the units changes the covariance. A unit-free measure of the relationship between X and Y is the **correlation**.[58]

[58] The usual Greek letter representing correlation is ρ.

The correlation is 1 if there is a perfect positive *linear* relationship between X and Y, -1 if there is a perfect negative *linear* relationship, and 0 if there is no *linear* relationship between X and Y.[59] Thus $|\rho|$ determines the *strength* of the relationship[60] between X and Y and $\text{sign}(\rho)$ determines the *direction* of the relationship.[61]

[59] Notice the emphasis on the word "linear"; there can be a relationship between X and Y that would make their correlation small yet there could still be a strong nonlinear relationship linking the two variables.

[60] We can classify the strength of the relationship between rvs using completely arbitrary cutoffs; specifically, we could say that if $|\rho| < 0.3$ there is no notable correlation, if $|\rho| > 0.7$ there is a strong correlation, and otherwise the correlation is weak.

[61] There is a sample statistic for estimating ρ from paired data (x_1, y_i):

$$r = \frac{1}{s_x s_y (n-1)} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

The interpretation is the same. We do not discuss the sample statistic in this course.

Example 6

Compute the covariance and correlation for the random variables mentioned in Examples 1 and 2.

```
## For Example 1
(sigma_xy <- E(X * Y) - E(X) * E(Y))

## [1] 0
```

If X and Y are independent, then $\text{Cov}(X, Y) = 0$.[62] The converse is *not* true in general, as the following example shows.

[62] As a consequence, we can equivalently say that $\mathbb{E}[XY] = \mathbb{E}[X]\,\mathbb{E}[Y]$ when X and Y are idependent.

Example 7

The point (U, V) is equally likely to be any of the points in the sample space $\mathcal{S} = \{(1,1), (1,-1), (-1,2), (-1,-2)\}$. Compute $\text{Cov}(U, V)$. Are U and V independent?

```
(UV <- jointRV(list(c(-1, 1), c(-2, -1, 1, 2)),
               probs = c(1/4, 0, 0, 1/4, 0, 1/4, 1/4, 0)))
```

```
## Random variable with 4 outcomes
##
## Outcomes -1,-2  -1,2  1,-1   1,1
## Probs      1/4   1/4   1/4   1/4
```

```
(U <- marginal(UV, 1))
```

```
## Random variable with 2 outcomes
##
## Outcomes  -1   1
## Probs    1/2 1/2
```

```
(V <- marginal(UV, 2))
```

```
## Random variable with 4 outcomes
##
## Outcomes  -2  -1   1   2
## Probs    1/4 1/4 1/4 1/4
```

```
E(U*V) - E(U) * E(V)
```

```
## [1] 0
```

```
independent(U, V)
```

```
## [1] FALSE
```

We say (X_1, X_2) follows the bivariate Normal distribution, or $(X_1, X_2) \sim \text{BINORM}(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho)$, if the joint pdf of X_1 and X_2 is

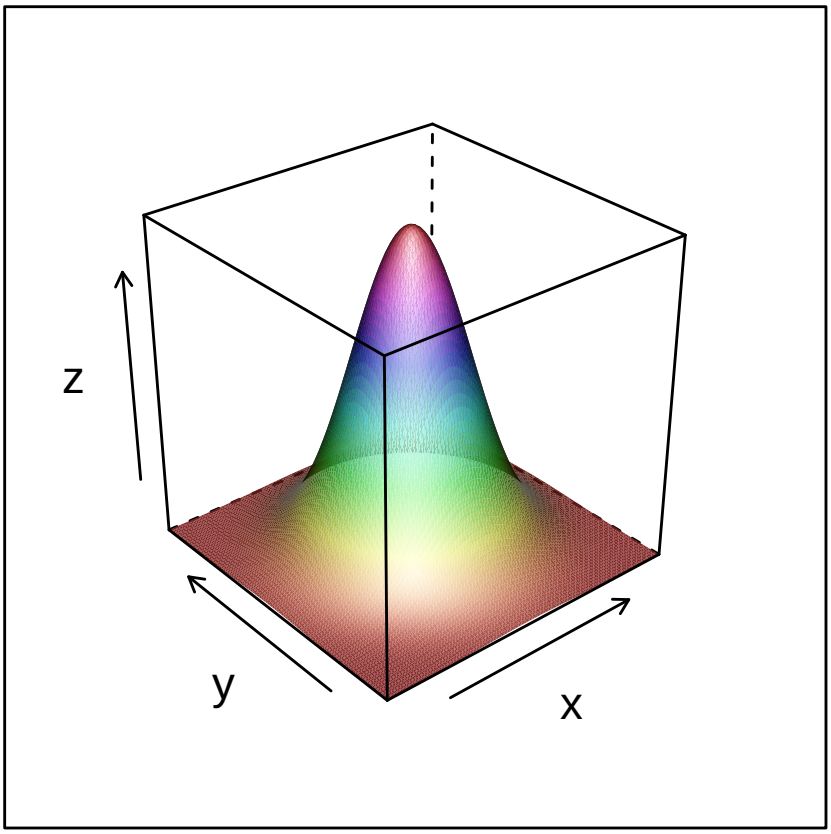
The pdf of the bivariate Normal distribution is illustrated below.

```
library(mvtnorm)
library(lattice)

my.settings <- list(superpose.polygon = list(border = "transparent"))
```

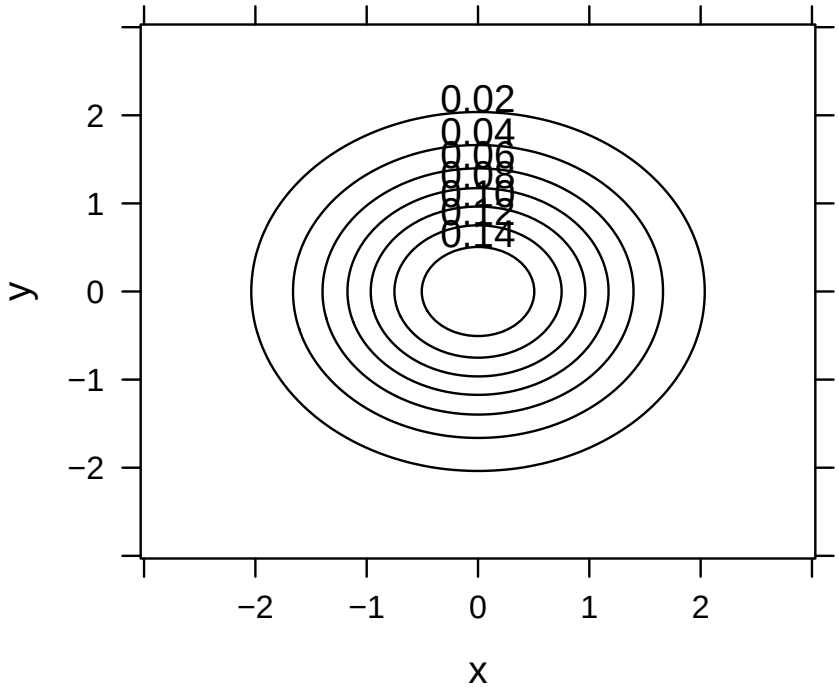
```
points <- data.frame("x" = rep(seq(-3, 3, length.out = 100), times = 100),
                     "y" = rep(seq(-3, 3, length.out = 100), each = 100))
points$z <- apply(points, 1, function(r) {dmvnorm(r)})
head(points)
```

```
##           x  y           z
## 1 -3.000000 -3 1.964128e-05
## 2 -2.939394 -3 2.351445e-05
## 3 -2.878788 -3 2.804818e-05
## 4 -2.818182 -3 3.333337e-05
## 5 -2.757576 -3 3.946923e-05
## 6 -2.696970 -3 4.656321e-05
```

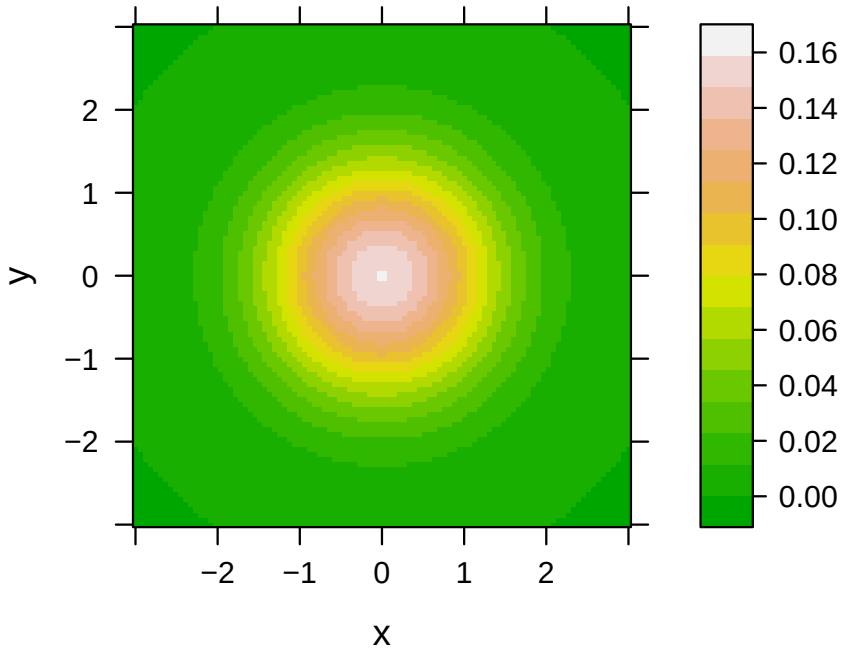
```
wireframe(z ~ x * y, data = points, lines = FALSE,
          col = "transparent", shade = TRUE)
```




```
contourplot(z ~ x * y, data = points)
```




```
levelplot(z ~ x * y, data = points, drape = TRUE,
          col.regions = terrain.colors(100))
```



If you were to slice the pdf in any direction, the resulting plot would be another Normal distribution. Specifically, the marginal distributions $f_{X_1}(x_1)$ and $f_{X_2}(x_2)$ are Normal distributions, and conditional distributions $f_{X_1|X_2}(x_1|x_2)$ and $f_{X_2|X_1}(x_2|x_1)$ are all all Normal distributions.

We have $\mathbb{E}[X_1] = \mu_1$, $\mathbb{E}[X_2] = \mu_2$, $\text{SD}(X_1) = \sigma_1$, $\text{SD}(X_2) = \sigma_2$, and $\text{Corr}(X_1, X_2) = \rho$. Crucially, when (X_1, X_2) follows a bivariate Normal distribution, $\text{Cov}(X_1, X_2)$ *does* imply independence![63]

[63] This is not the same as saying that if two random variables are Normally distributed and uncorrelated they are independent. Joint normality does not follow from the normality of the marginal distributions; for example, if we choose $Z_1 \sim N(0,1)$ and $Z_2 = SZ_1$ with $\mathbb{P}(S = 1) = \mathbb{P}(S = -1) = \frac{1}{2}$, then $Z_2 \sim N(0,1)$, and the marginal distributions of (Z_1, Z_2) are thus standard Normal distributions, and $\text{Cov}(Z_1, Z_2) = 0$. However, Z_1 and Z_2 are obviously *not* independent since if we know Z_1 then we know Z_2 differs from Z_1 by at most a sign.

Example 8

Let H_C represent the height of a son and H_F the height of the son's father (in inches). Suppose

$$(H_C, H_F) \sim \text{BINORM}(69.2, 69.2, 2.6, 2.6, 0.4)$$

1. What are the marginal distributions of H_C and H_F?

2. Suppose a person's father is 78 inches tall. Find an equal-tailed[64] interval such that the probability the child's height is in this interval is 0.95.

[64] In general, when we say an interval is equal-tailed, we mean that the probability that the random variable is too small to be in the region is equal to the probability that the random variable is too large. We need this restriction in order to have a unique solution; otherwise, there could be an infinite number of solutions.

```r
## Marginal distributions are trivial; let's worry about the conditional
mu1 <- 69.2; mu2 <- 69.2; sigma1 <- 2.6; sigma2 <- 2.6; rho <- 0.4
h <- 78
(mu_2g1 <- (mu1 - rho * mu2 * sigma1 / sigma2) + rho * sigma1 * h / sigma2)
```

```
## [1] 72.72
```

```r
(sigma_2g1 <- sqrt((1 - rho^2) * sigma1^2))
```

```
## [1] 2.382939
```

```r
qnorm(0.025, mean = mu_2g1, sd = sigma_2g1)  # Lower bound
```

```
## [1] 68.04952
```

```r
qnorm(0.975, mean = mu_2g1, sd = sigma_2g1)  # Upper bound
```

```
## [1] 77.39048
```

Section 5: The Distribution of a Linear Combination[65]

[65] In my opinion, Section 5 of this chapter is a more logical successor of Section 2; we will come back to Section 3 later.

Consider a collection of n random variables $X_1, \ldots, X_n$ and numerical constants $a_1, \ldots, a_n$. The rv Y is a **linear combination** of the random variables $X_1, \ldots, X_n$ if Y is of the form

Proposition 13. *Suppose* $\mathbb{E}[X_i] = \mu_i$ *and* $\text{Var}(X_i) = \sigma_i^2$. *The following facts are true:*

Thus we have the important property about expectations; they are linear operators, to use linear algebra language. The variance is not a linear operator (although it is a sublinear operator), but the covariance is a bilinear operator (linear in both its arguments).

Corollary 1. *Suppose X_1 and X_2 are two independent random variables. Then:*

We can weaken the independence assumption to simply being uncorrelated and the variance computation will still be true.

Proposition 14. *Suppose X_1 and X_2 are two Normal random variables. Then*

Corollary 2. *A linear combination of Normal random variables also follows a Normal distribution.*

Example 9

Suppose $X_1, \ldots, X_n$ are i.i.d. random variables. Compute the expected value, variance, and standard deviation of $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

Example 10

Suppose $X_1, \ldots, X_n$ are i.i.d. random variables. Compute the expected value, variance, and standard deviation of $T = \sum_{i=1}^{n} X_i$

There are several random variables where we know the distribution of sums of those random variables. Below is a summary:

Section 3: Statistics and Their Distributions

We will call the collection $X_1, \ldots, X_n$ a **random sample** if it consists of i.i.d. random variables. We will call any quantity we can compute from a random sample a **statistic**. Before the dataset is observed, a statistic is a random quantity, with its own distribution, referred to as the **sampling distribution**; statistics in this random state are usually referred to using upper-case letters, while the observed statistic (after we have a dataset) is usually referred to using lower-case letters.

Examples of statistics include:

Example 11

Let $X_1, \ldots, X_n$ be i.i.d.r.v. with $X_1 \sim \text{Ber}(p)$. What is the sampling distribution of $\bar{X}$?

Example 12

Let $X_1, \ldots, X_n$ be i.i.d.r.v. with $X_1 \sim N(\mu, \sigma)$. What is the sampling distribution of $\bar{X}$? Use the sampling distribution to find an interval such that $\mathbb{P}\left(l(\bar{X}) \leq \mu \leq u(\bar{X})\right) = 1 - \alpha$.

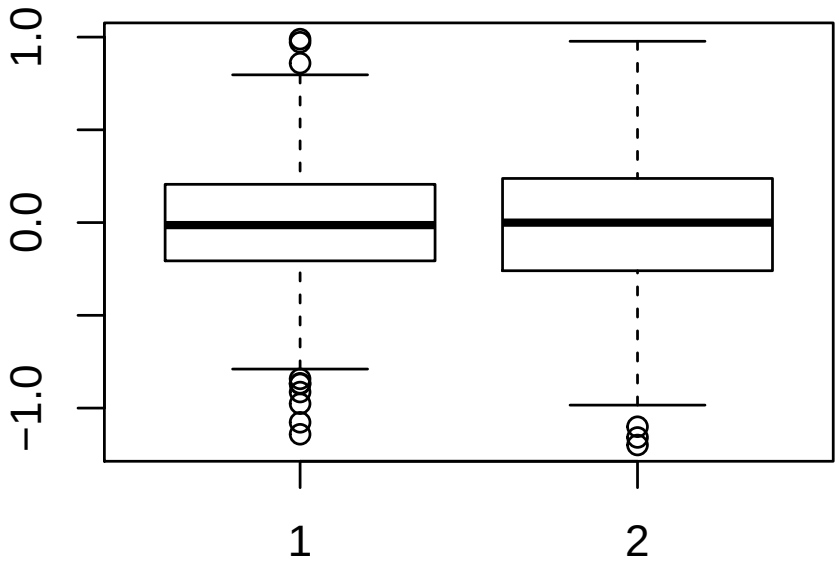
One approach to finding information about the sampling distribution of statistics is to use **simulations**. We generate K random samples of size n, $X_{1,k}, \ldots, X_{n,k}$, $k \in [K]$. For each sample we compute the statistic of interest $T(X_{1,k}, \ldots, X_{n,k}) = T_k$, and then study the random sample $T_1, \ldots, T_K$.

Example 13

For the Normal distribution, we could estimate the parameter μ using either the sample mean $\bar{X}$ or the sample median $\tilde{X}$. What are the properties of these two statistics' sampling distributions? What are their respective shapes? Which has a smaller variance?

Let's suppose $X_1 \sim N(0,1)$, then conduct a simulation study to compare these statistics. We'll look at $n = 10$ and use $K = 1000$ samples.

```
## Generate 1000 random samples of size ten, storing them in a 10x1000 matrix
datamat <- replicate(1000, rnorm(10))
sim_mean <- apply(datamat, 2, mean)
sim_med <- apply(datamat, 2, median)
boxplot(sim_mean, sim_med)
```




```
summary(sim_mean)
```

```
##      Min.   1st Qu.    Median      Mean
## -1.141278 -0.206299 -0.013537 -0.008187
##   3rd Qu.      Max.
##  0.205722  0.988908
```

```
summary(sim_med)
```

```
##       Min.    1st Qu.     Median       Mean
## -1.1992114 -0.2592222 -0.0004545 -0.0110169
##    3rd Qu.       Max.
##  0.2374234  0.9773014
```

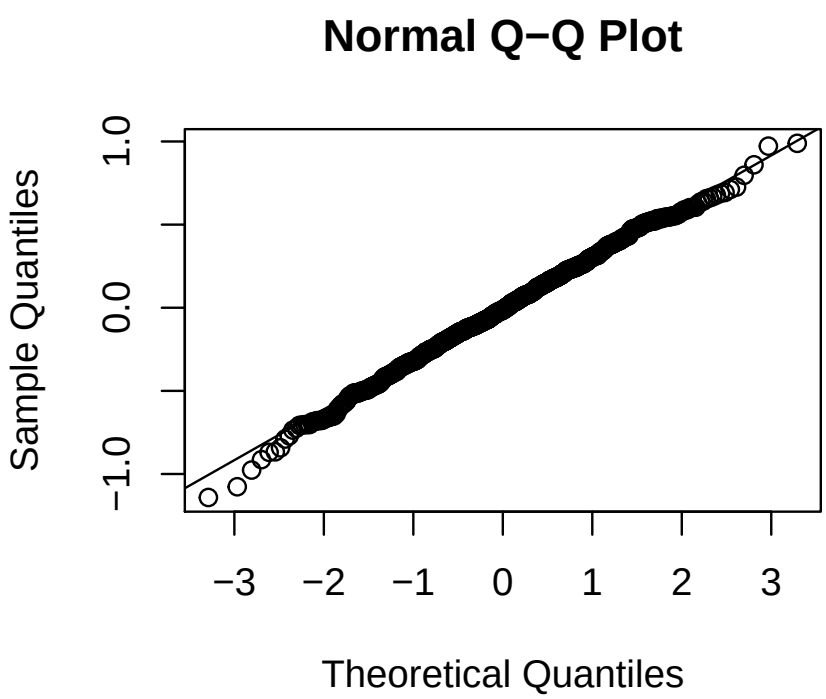
```
var(sim_mean)
```

```
## [1] 0.0983803
```

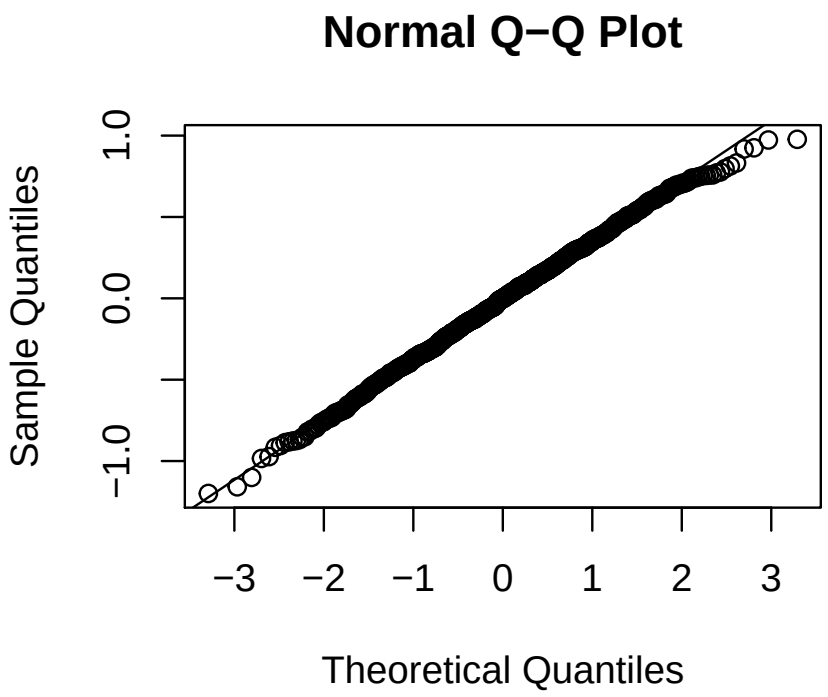
```
var(sim_med)
```

```
## [1] 0.1318241
```

```
qqnorm(sim_mean); qqline(sim_mean)
```




```
qqnorm(sim_med); qqline(sim_med)
```



Example 14

Let's now consider the sample mean of random samples $U_1, \ldots, U_n$ with $U_1 \sim \text{UNIF}(0,1)$. What can we say about the distribution of the sample mean $\bar{U}$ as the sample size n gets large?

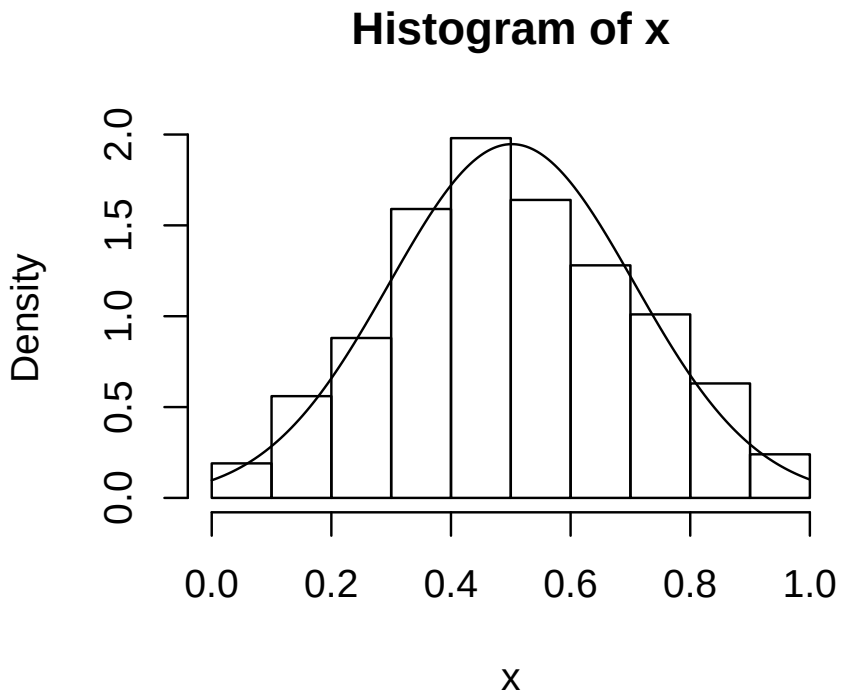
We will create 1000 samples for $n \in \{5, 20, 80\}$, then compare the distributions.

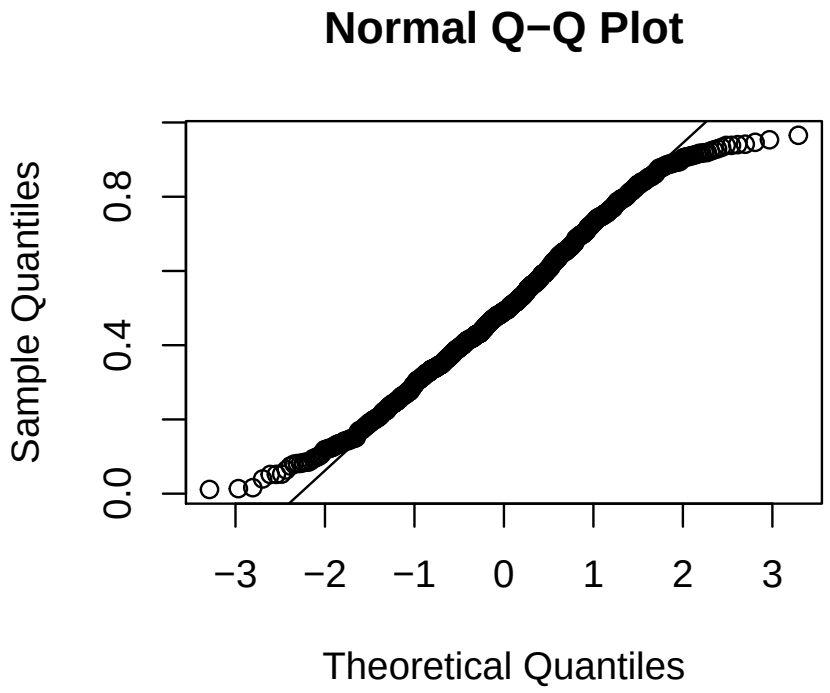
```
sizes <- c(2, 5, 20, 80)
k <- 1000

datasets <- lapply(sizes, function(n) {
  replicate(k, runif(n))
})
names(datasets) <- sizes
str(datasets)

## List of 4
##  $ 2 : num [1:2, 1:1000] 0.575 0.135 0.125 0.898 0.284 ...
##  $ 5 : num [1:5, 1:1000] 0.294 0.179 0.893 0.965 0.32 ...
##  $ 20: num [1:20, 1:1000] 0.418 0.712 0.339 0.192 0.19 ...
##  $ 80: num [1:80, 1:1000] 0.98088 0.74066 0.00726 0.35442 0.08737 ...

sim_mean_unif <- lapply(datasets, function(d) {apply(d, 2, mean)})
str(sim_mean_unif)
```

```
## List of 4
##  $ 2 : num [1:1000] 0.355 0.511 0.632 0.475 0.345 ...
##  $ 5 : num [1:1000] 0.53 0.419 0.656 0.573 0.533 ...
##  $ 20: num [1:1000] 0.487 0.468 0.416 0.47 0.603 ...
##  $ 80: num [1:1000] 0.416 0.51 0.434 0.496 0.508 ...
```

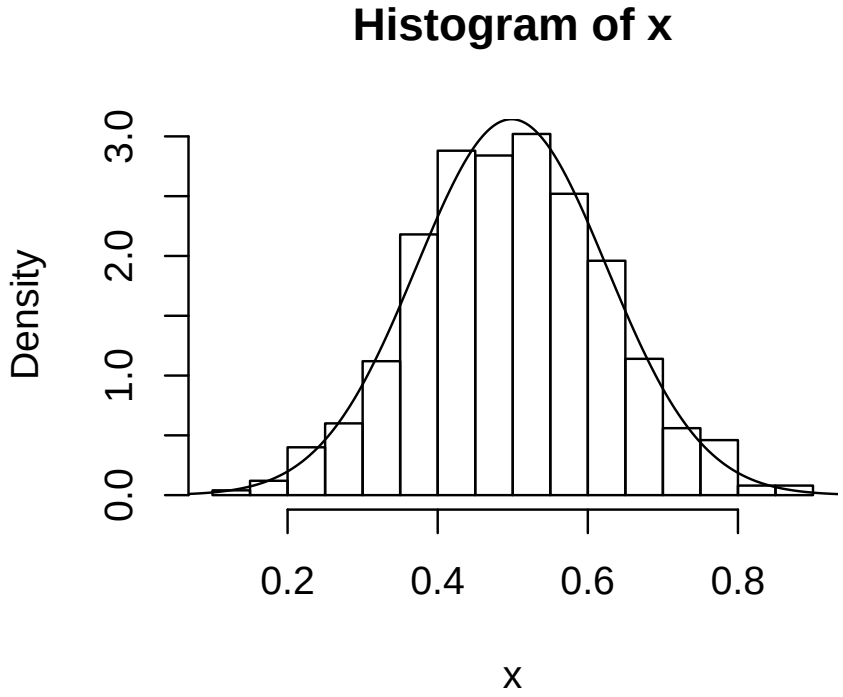
```
for (x in sim_mean_unif) {
  print(summary(x))
  hist(x, freq = FALSE)
  lines(seq(0, 1, length.out = 1000),
        dnorm(seq(0, 1, length.out = 1000), mean = mean(x), sd = sd(x)))
  qqnorm(x); qqline(x)
}
```

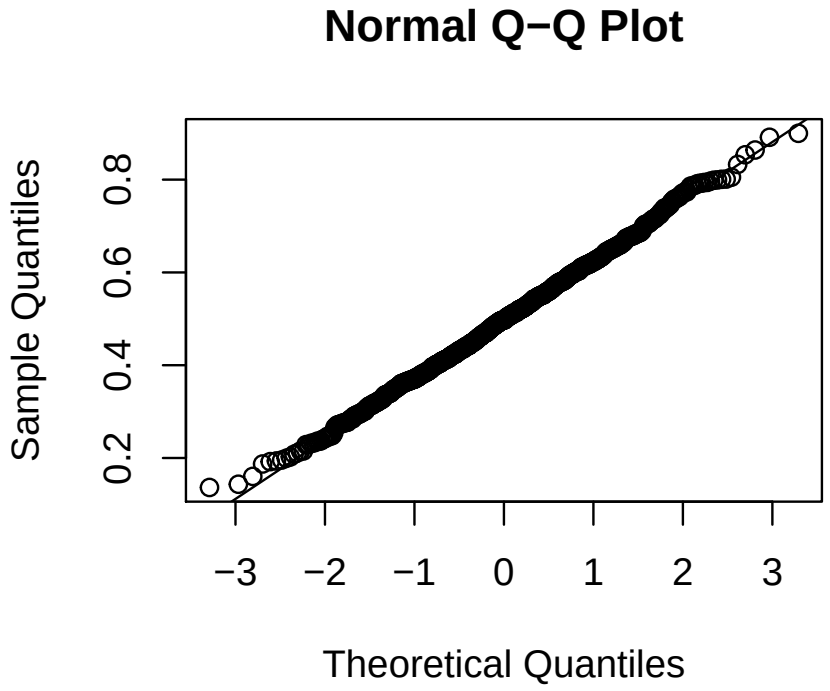
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.
## 0.01131 0.35465 0.49095 0.50179 0.65230
##    Max.
## 0.96549
```



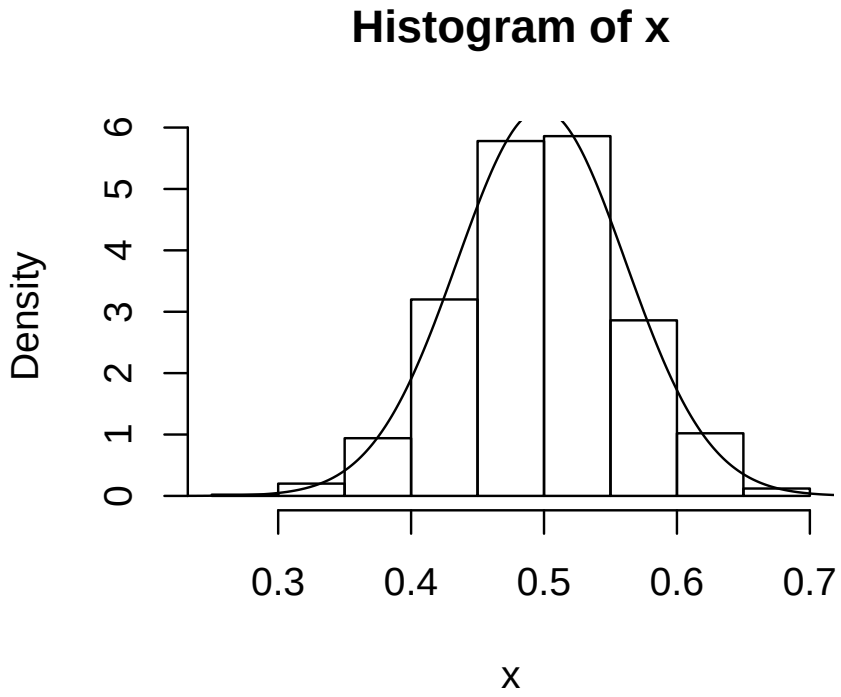


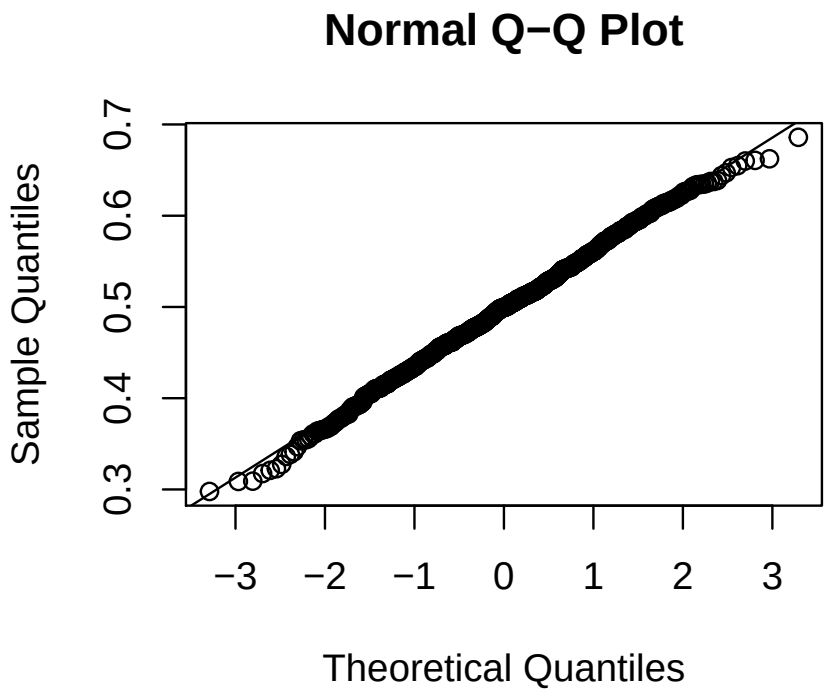

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.
##  0.1365  0.4100  0.4975  0.4984  0.5829
##    Max.
##  0.8999
```



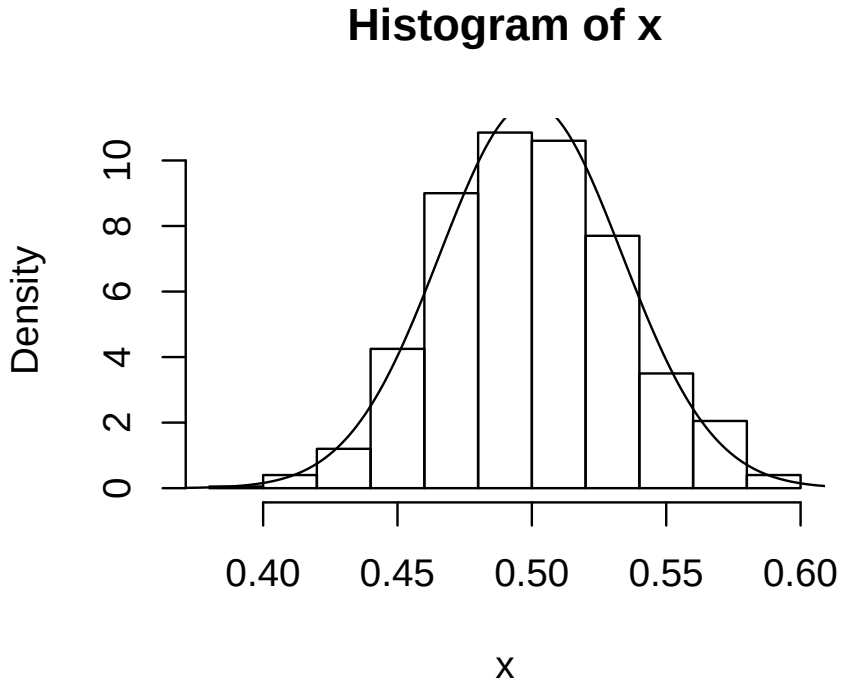


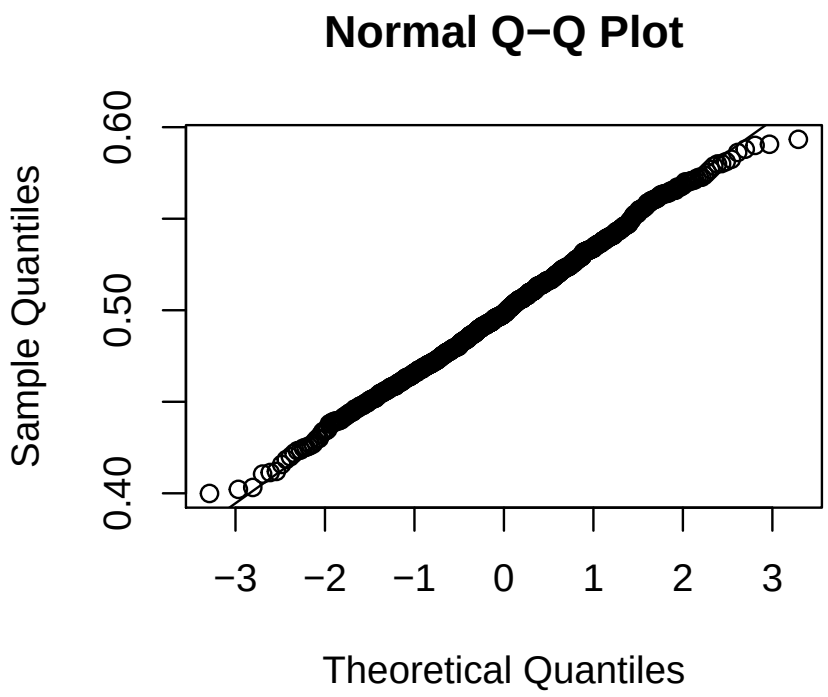

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.
##  0.2978  0.4574  0.4995  0.4980  0.5412
##    Max.
##  0.6859
```






```
##    Min. 1st Qu.  Median    Mean 3rd Qu.
##  0.3999  0.4756  0.4981  0.4996  0.5226
##    Max.
##  0.5934
```





Section 4: The Distribution of the Sample Mean

There are two theorems that form the cornerstone of probability and statistics: the law of large numbers and the central limit theorem. The Law of Large Numbers (LLN) guarantees us that the sample mean will approximately equal the population mean, while the Central Limit Theorem (CLT) describes the distribution of the sample mean for large n when we have a mean and a variance.

Theorem 3 (Law of Large Numbers)**.** *Let $X_1, X_2, \ldots$ be a sequence of of i.i.d.r.v. with $\mathbb{E}[X_1] = \mu$, and let $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. Then the probability the (random) sequence $\bar{X}_n$ converges to μ is 1.*

Theorem 4 (Central Limit Theorem)**.** *Under the same assumptions as the Law of Large Numbers but with the additional assumption that* $\text{Var}(X_1) = \sigma^2$, $\mathbb{P}\left(\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \leq z\right) \to \Phi(z)$ *for all z. In other words, the distribution of $\bar{X}_n$ is approximately $N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$, with the approximation improving as $n \to \infty$.*

More directly, the CLT describes the behavior of sums of i.i.d.r.v. as more random variables are summed. The CLT explains why some distributions–like the binomial distribution, the Poisson distribution, the gamma distribution, and the χ^2-distribution–can be approximated with the Normal distribution as one of their parameters grows

large; these distributions can be interpreted as the distributions of sums of i.i.d.r.v.[66] and thus the CLT applies.[67]

Thanks to the CLT, we can describe the distribution of the sample mean without worrying about the exact distribution of the underlying data if the sample size n is large enouch[68], since the CLT says that the initial distribution is eventually "forgotten" by the sample mean.

[66] Specifically: $\text{BIN}(n,p)$ is the sum of n $\text{Ber}(p)$ r.v.s; $\text{POI}(n)$ is the sum of n $\text{POI}(1)$ r.v.s; $\text{GAMMA}(\alpha,n)$ is the sum of n $\text{EXP}(\alpha)$ r.v.s; and $\chi^2(n)$ is the sum of n $\chi^2(1)$ r.v.s.

[67] The CLT requires that $\text{Var}(X_1) < \infty$; if this does not hold, the CLT no longer applies and its conclusion may not even be true.

[68] In general it's safe to use the CLT if $n > 30$.

Example 15

The average customer visiting a grocery store spends X dollars, where $\mathbb{E}[X] = 50$ and $\text{SD}(X) = 55$.[69] Every month about 30,000 purchases are made at the grocery store.

[69] Notice X is non-negative but $\text{SD}(X) > \mathbb{E}[X]$. This can happen with skewed distributions.

1. What will be the (approximate) distribution of the average purchase, $\bar{X}$?

2. What is the (approximate) probability that the revenue of the grocery store in a month is less than \$1,485,000

```
55/sqrt(30000)  # sd of xbar
```

```
## [1] 0.3175426
```

```
pnorm(1485000, mean = 50 * 30000, sd = 55 * sqrt(30000))
```

```
## [1] 0.05767537
```

Chapter 6: Point Estimation

Introduction

Karl Pearson, perhaps the first mathematical statistician, proposed the modern view that the objective of science is to estimate the parameters of a probability distribution that generates datasets (Salsburg, 2002). Statistics has come a long way since Karl Pearson's methods, and in this chapter (where we finally leave our study of probability behind to dive into statistics) we see how to compute estimates for distribution parameters.

Initially there are many statistics competing to estimate some quantity; for example, both the sample mean and sample median could estimate the parameter μ of the Normal distribution. In the first section, we see general principles used to evaluate estimators. In the second section, we see methods for generating estimates.

Section 1: Some General Concepts of Point Estimation

There are many parameters we may try to estimate, such as

- μ from the distribution EXP(μ)
- μ and σ from the Normal distribution $N(\mu, \sigma)$
- α and β from the Weibull distribution WEI(α, β)
- And others

We want to discuss parameters and estimators using a general language. Let θ be a parameter, and $\hat{\theta}$ is an estimator for θ. Often the notation $\hat{\theta}$ refers to both a random variable and a specific point estimate.[70] We call $\hat{\theta}$ a **point estimator** for θ; we use the point estimator to compute a **point estimate**, a single plausible value for θ.

Examples of point estimators and the parameters they estimate include:

[70] I've said that usually capital letters refer to random variables; in this case, we would use $\hat{\Theta}$ to refer to the random version of the estimator, and $\hat{\theta}$ to refer to a specific number computed from an observed, no-longer-random dataset. However, this is not conventional; writers are lazy and don't like writing $\hat{\Theta}$, preferring $\hat{\theta}$ instead. Readers can usually tell whether the writer is referring to a random number or a computed number. As I said, capital letters *usually* refer to random variables; this is one of the (many) exceptions.

An estimator $\hat{\theta}$ is an **unbiased estimator** for θ if

---

Example 1

Show that the sample mean $\bar{X}$ computed from iid data is an unbiased estimator for the population mean μ.

Example 2

Suppose that $X_1, \ldots, X_n$ is an iid sample from a Bernoulli distribution with parameter p. Show that the sample proportion is an unbiased estimator for p.

Example 3

Show that the estimator $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$ computed from an iid sample $X_1, \ldots, X_n$ with $\text{Var}(X_1) = \sigma^2$ is *not* an unbiased estimator for σ^2.

Example 4

Show that the sample variance $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ computed from an iid sample $X_1, \ldots, X_n$ with $\text{Var}(X_1) = \sigma^2$ is an unbiased estimator for σ^2.[71]

[71] It's tempting to think that the sample standard deviation $S = \sqrt{S^2}$ is an unbiased estimator for σ, but this is *not* the case; S is a biased estimator for σ, with a tendency to underestimate the true σ. However, S is justified by other criteria. In fact, estimation of σ presents a good case study in why unbiasedness, as a criterion for good estimators, may be overrated (see Wikipedia (2018)).

Example 5

Suppose $X_1, \ldots, X_n$ is an iid sample from an exponential distribution with mean μ. Recall that the rate parameter of an exponential distribution is $\lambda = \frac{1}{\mu}$. Show that the estimator $\hat{\lambda} = \frac{1}{\bar{X}}$ is *not* an unbiased estimator for μ.

Suppose we want to estimate μ for Normally distributed data. $\bar{X}$ is an unbiased estimator for μ. So is $\tilde{X}$. In fact, X_1 is an unbiased estimator for μ since $\mathbb{E}[X_1] = \mu$. The last estimator is clearly silly, but not because of the unbiasedness criterion. Instead, the last estimator violates the **minimum variance** criterion, which states that the standard error (the standard deviation of $\hat{\theta}$, referred to as $\sigma_{\hat{\theta}}$) should be as small as possible, if not the smallest of all possible estimators. In this case, of the estimators I just mentioned, $\bar{X}$ has the smallest variance, and X_1 the largest. In fact, $\bar{X}$ is the **minimum variance unbiased estimator (MVUE)** for μ in this context, having the smallest variance of any unbiased estimator of μ. Likewise, $\hat{p}$ is the MVUE for p, when the data was drawn from a Bernoulli distribution with parameter p.

The minimum variance and unbiasedness criteria are not necessarily in agreement; there may be an estimator that has a smaller variance than all unbiased estimators and is close to the true value of θ when sample sizes are large. We may relax the unbiasedness criterion and instead require **consistency**, which says that a law of large numbers applies to the estimator; that is, $\hat{\theta}_n \to \theta$ in some sense as n grows (with $\hat{\theta}_n$ being an estimator for θ computed from n data points). The only estimator mentioned so far that isn't consistent is X_1; the rest (including the sample standard deviation) are consistent estimators.

Sometimes an estimator performs well in some circumstances but poorly in others; for example, $\bar{X}$ estimates the location of a distribution well when data is drawn from a Normal distribution but poorly when computed from data drawn from a distribution with heavy tails, such as the Laplace or Cauchy distributions. We call an estimator **robust** when the estimator performs well in multiple scenarios. Trimmed means, for example, as seen as robust estimators for the location of a distribution.

The **standard error** of an estimator is defined below:

The standard error can depend on unknown parameters. In that case, we may report an **estimated standard error**, where estimates for the unknown parameters are used in those parameters' place. Estimates of standard errors are often reported with point estimates to give a sense of how accurate the point estimate is. We will see how standard errors are often used to compute plausible regions for the location of θ in Chapter 7.

Example 6

Suppose $\text{Var}(X_1) = \sigma^2$ and the dataset $X_1, \ldots, X_n$ is an iid dataset. What is the standard error of $\bar{X}$? Use this to give estimates of standard errors for data drawn from Normal, exponential, and Poisson distributions.

Example 7

Suppose $X_1, \ldots, X_n$ is a Bernoulli dataset. What is the standard error of $\hat{p}$? What is an upper bound on the standard error? What is an estimate of the standard error?

Bootstrapping is a computer intensive technique for computing the standard errors of estimates. Bootstrap estimates of standard errors are often robust and allow us to obtain estimates when formulas for those errors would be intractable.

Suppose that $x_1, \ldots, x_n$ is a sample of iid data drawn from a distribution with pdf $f(x;\theta)$. The bootstrap procedure works as follows:

1. Estimate θ with $\hat{\theta}$.
2. Choose a large number B.
3. Generate B samples of data $X_{b1}, \ldots, X_{bn}$ ($1 \leq b \leq B$) from the distribution with pdf $f(x;\hat{\theta})$, and from each of them compute $\hat{\theta}_b^*$, the estimate of θ using $x_{b1}, \ldots, x_{bn}$; you should now have a collection of data $\hat{\theta}_1^*, \ldots, \hat{\theta}_B^*$
4. Compute $\bar{\theta}^* = \frac{1}{B}\sum_{b=1}^{B} \hat{\theta}_b^*$; this is the bootstrap estimate of θ
5. Compute $\hat{\sigma}_{\hat{\theta}} = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\theta}_b^* - \bar{\theta}^*\right)^2}$; this is the bootstrap standard error estimate

---

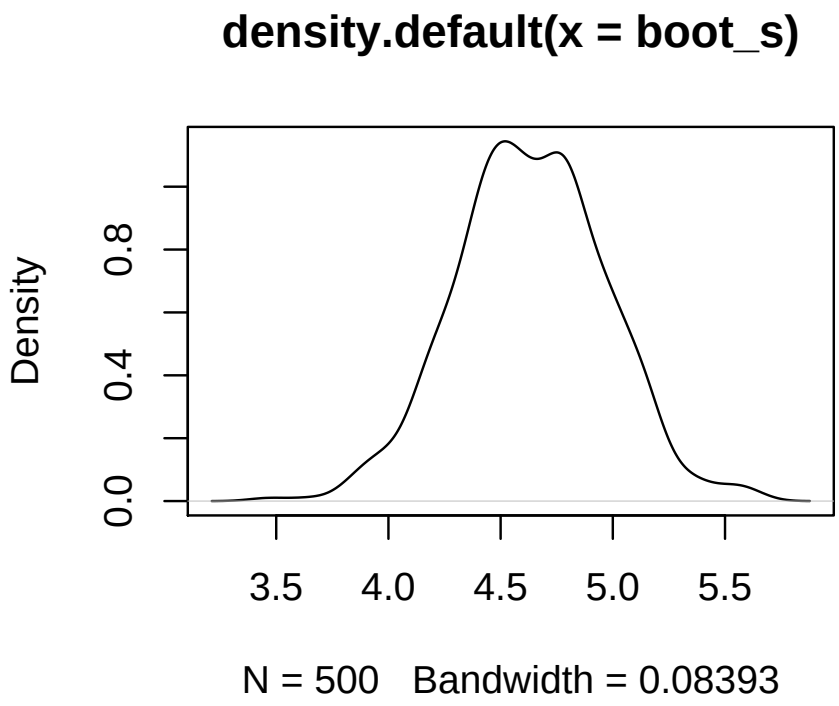
Example 8

In this example I demonstrate how to estimate the standard error of the estimate of the sample standard deviation computed from Normally distributed data.

```
n <- 100  # Our sample size is 100
B <- 500  # The bootstrap sample size is 500
dat <- rnorm(n, mean = 10, sd = 5)  # Our dataset
(s <- sd(dat))  # Estimated standard deviation
```

```
## [1] 4.647059
```

```
boot_s <- replicate(B, {
  boot_dat <- rnorm(n, mean(dat), s)
  sd(boot_dat)
})
plot(density(boot_s))
```




```
mean(boot_s)  # The bootstrap estimator of s
```

```
## [1] 4.635094
```

```
sd(boot_s)  # The bootstrap-estimated standard error of s
```

```
## [1] 0.3317313
```

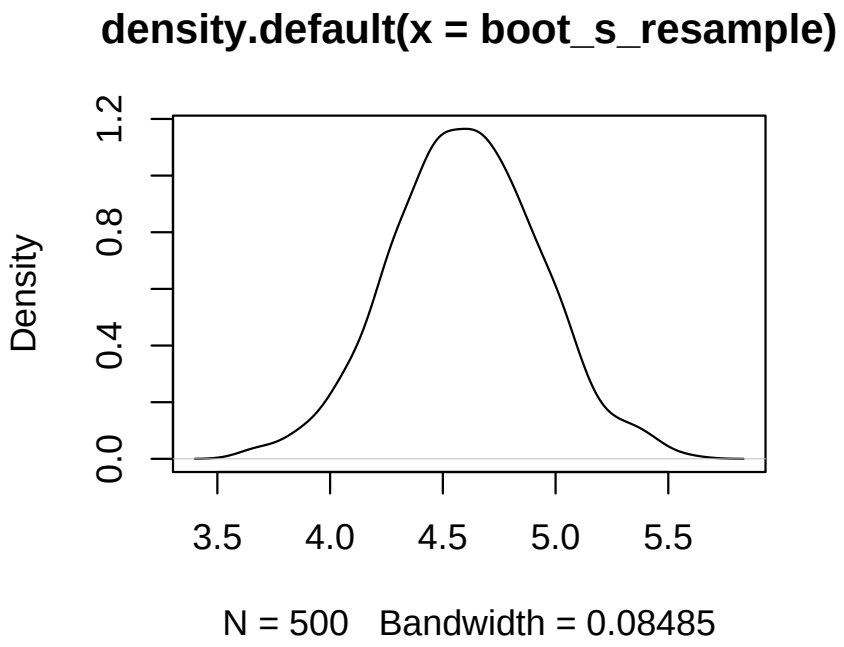
If we don't want to assume that the data came from a particular sample, we can sample instead from the data itself, doing so with replacement. When doing this, we are said to be sampling from the empirical cdf, or empirical distribution, of the data; that is, we are sampling from the distribution we observed, which serves as an estimate of the population distribution that generated the data.

Example 9

This example demonstrates obtaining a bootstrap estimate of the standard error of the standard deviation without assuming that the data was drawn from a particular distribution, using the resampling technique.

```
boot_s_resample <- replicate(B, {
  boot_dat <- sample(dat, n, replace = TRUE)
```

```
  sd(boot_dat)
})
plot(density(boot_s_resample))
```




```
mean(boot_s_resample)
```

```
## [1] 4.600309
```

```
sd(boot_s_resample)
```

```
## [1] 0.3278774
```

Section 2: Methods of Point Estimation

Assume again that $X_1, \ldots, X_n$ is an iid sample from some distribution. $\mathbb{E}\left[X_1^k\right]$ is called the k^{th} **population moment**, and $\frac{1}{n}\sum_{i=1}^n X_i^k$ is called the k^{th} **sample moment**. As an example, $\bar{X}$ is the first sample moment and $\mathbb{E}[X_1] = \mu$ is the first population moment.[72] A sample moment is an unbiased estimator for the corresponding population moment.

[72] σ^2 is related to the second sample moment but isn't the second moment itself. The same goes for S^2 and sample moments.

Suppose a distribution has $\theta_1, \ldots, \theta_K$ parameters we wish to estimate. Below is the **method of moments estimation** procedure:

1. Compute the first K population moments, $m_1, \ldots, m_K$ in terms of the unknown parameters $\theta_1, \ldots, \theta_K$
2. Solve for $\theta_1, \ldots, \theta_K$ so they are expressed in terms of $m_1, \ldots, m_K$

3. Replace $m_1, \ldots, m_K$ with $M_1, \ldots, M_K$, the first K sample moments; the resulting expressions are $\hat{\theta}_1, \ldots, \hat{\theta}_K$, the **method of moments estimators (MMEs)** for the desired parameters.

Method of moments estimation produces consistent estimators for desired parameters using an intuitive procedure. There is no guarantee the estimators are unbiased (in fact they likely are not unbiased) and they usually are not minimum-variance estimators. In fact, in the context the estimators were computed, there likely is an estimator that is consistent and with a smaller variance than the MMEs That said, method of moment estimators are often robust and more tractable than other estimators while being easy to compute.[73]

[73] Method of moments estimation is often used in economics due to their simplicity and robustness.

Example 10

What is the method of moments estimator for the population variance?

Example 11

What is the MME for the rate parameter $\lambda = \frac{1}{\mu}$ for an exponential distribution?

Example 12

Let $X_1, \ldots, X_n$ be an iid sample from the distribution UNIF$(a - b, a + b)$. What are the MMEs for a and b?

Example 13

Consider a shifted exponential distribution that depends on two parameters μ and γ such that $X_1 - \gamma \sim \text{EXP}(\mu)$. What are the MMEs for μ and γ?

To illustrate the principle of the next estimation method, suppose I flip a coin and record whether I get heads or not. The coin could be a fair coin or a biased coin, where the probability of getting heads is $p = .9$. When I flip the coin and observe an outcome, how will I decide which coin was flipped?

Consider the following table:

After flipping the coin and observing the outcome, I look to the table to see what the probability of that outcome was under each scenario of coin choice. The maximum likelihood principle says that I should choose the coin that maximizes these probabilities.

Let $X_1, \ldots, X_n$ have the joint pmf/pdf $f(x_1, \ldots, x_n; \theta_1, \ldots, \theta_K)$. When $x_1, \ldots, x_n$ are the observed values of the dataset, this function is called the **likelihood function** when it is regarded as a function of $\theta_1, \ldots, \theta_K$, as expressed below:

When the random variables $X_1, \ldots, X_n$ are iid, the likelihood function is

The **maximum likelihood estimators (MLEs)** $\hat{\theta}_1, \ldots, \hat{\theta}_K$ are the values that maximize the likelihood function. They are interpreted as the most likely values of the parameters given the data we saw, in that we were most likely to see the values of the data if those were the parameters.

Usually the likelihood function is hard to maximize on its own, so instead we maximize the log-likelihood function

Since $\ln(x)$ is an increasing function, both functions have the same

maxima.

Example 14

Consider an iid dataset of Bernoulli data. What is the maximum likelihood estimator of the sample proportion p?

Example 15

Consider an iid dataset drawn from the EXP(μ) distribution. Find the MLE for μ.

Example 17

Consider an iid dataset drawn from the $N(\mu, \sigma^2)$ distribution. Find the MLE of μ and σ^2.

Example 18

Consider an iid dataset drawn from the UNIF$(0, \theta)$ distribution. Find the MLE of θ.

MLEs are consistent estimators and are either minimum variance or almost minimum variance, with these properties improving as the sample size grows large. Additionally, the MLE of a function of parameters $h(\theta_1, \ldots, \theta_K)$ is the value of that function when applied to the MLEs $h(\hat{\theta}_1, \ldots, \hat{\theta}_K)$.

Example 19

Expanding on Example 15, find the MLE of the rate parameter $\lambda = \frac{1}{\mu}$ of an exponential distribution.

Example 20

Expanding on Example 20, find the MLE of the standard deviation σ of a Normal distribution.

Maximum likelihood estimation is an example of a general approach to parameter estimation, where a "good" estimate is an estimate that optimizes some objective function. MLEs maximize the likelihood function. Least-squares estimators minimizes the sum of square errors, $\sum_{i=1}^{n} \left(x_i - \hat{x}_i(\hat{\theta}_1, \ldots, \hat{\theta}_K)\right)^2$ (with $\hat{x}_i(\hat{\theta}_1, \ldots, \hat{\theta}_K)$ being the predicted value of x_i based on the parameter estimates), and least absolute deviation estimators minimize $\sum_{i=1}^{n} \left|x_i - \hat{x}_i(\hat{\theta}_1, \ldots, \hat{\theta}_K)\right|$. M-estimators maximize $\sum_{i=1}^{n} \rho(x_i; \hat{\theta}_1, \ldots, \hat{\theta}_K)$, where the "objective function" ρ is chosen to give the resulting estimator desired robustness properties.

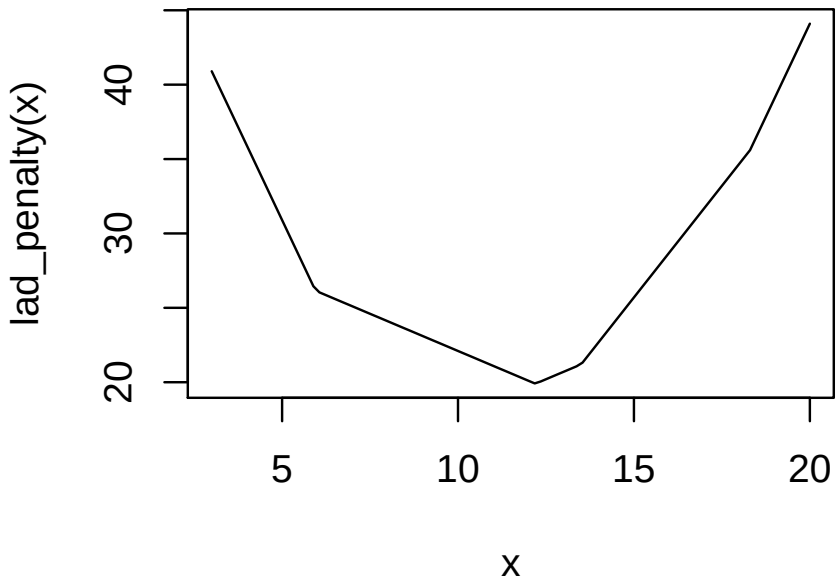
Example 21

Consider the following dataset:

```
x <- c(12.2, 18.3, 6.0, 5.9, 13.5)
```

The predicted value $\hat{\mu}$ of the data is the least absolute deviation estimator. Find the value of the estimator.

```
lad_penalty <- function(mu) {sum(abs(x - mu))}
lad_penalty <- Vectorize(lad_penalty)
curve(lad_penalty(x), 3, 20)
```



```
optim(0, lad_penalty)
```

```
## Warning in optim(0, lad_penalty): one-dimensional optimization by Nelder-Mead is unreliable:
## use "Brent" or optimize() directly
```

```
## $par
## [1] 12.3
##
## $value
## [1] 20
##
## $counts
## function gradient
##       24       NA
##
## $convergence
## [1] 0
##
## $message
## NULL
```

```
median(x)
```

```
## [1] 12.2
```

Chapter 7: Statistical Intervals Based on a Single Sample

Introduction

While we appreciate a parameter estimate we know that with any estimate there is uncertainty. Rather than report a single number, statisticians prefer to report a range of plausible values for the parameter being estimated. The shorter the range, the more we know about the location of the parameter.

In this chapter we will be looking at more common statistical intervals, such as confidence intervals. We will see how to construct them and how to properly interpret them. (Statisticians care a lot about the correct interpretation!)

Section 1: Basic Properties of Confidence Intervals

A $100(1-\alpha)$**% confidence interval (CI)** is a **random interval** (an interval with random endpoints) intended to describe the location of a parameter θ. Suppose the endpoints of the random interval are $l(x_1, \ldots, x_n)$ and $u(x_1, \ldots, x_n)$ (recall the distinction between x_i and X_i; here, the former is an observed number, perhaps from a sample, while X_i is a random variable). The CI for θ is an interval $(l(x_1, \ldots, x_n), u(x_1, \ldots, x_n))$ such that

$$\mathbb{P}\left(l(X_1, \ldots, X_n) \leq \theta \leq u(X_1, \ldots, X_n)\right) = 1 - \alpha$$

In short, in the long run, $100(1-\alpha)\%$ of intervals constructed this way capture the true value of θ.[74] Common confidence intervals include 90%, 95%, and 99%[75].

Suppose that σ is known and we have a dataset of i.i.d. data, with observed values $x_1, \ldots, x_n$. A confidence interval for the population mean μ is

[74] This is *not* the same as saying that the *probability* the interval captured θ is $1-\alpha$. The distinction is subtle but important. When we construct a confidence interval from a particular dataset, the endpoints are not random, and *that particular interval may or may not* include the true value of θ. We have to use this frequentist notion of a long-run capture rate in order to make sense of the interval. There are intervals out there where we can refer to the probability of whether a particular interval captured the true θ, such as the Bayesian **credible interval**, but this uses a completely different theory and interpretation of probability, in addition

This interval is exact when the data follows a Normal distribution[76] and approximately correct (due to the CLT) for large n when σ exists, for any underlying distribution. This interval takes the commonly-seen[77] form

[76] We got this interval in the Chapter 5 notes.

[77] This is not law; we will see intervals not of this form.

For this interval, the **margin of error (moe)** is

Consider for a second the variables involved in the margin of error, and consider changing their values. Which variables (all others being equal) lead to the margin of error being larger when they increase? Which would lead to a decrease in the margin of error?

Consider the denominator of the moe. What is the relationship between the amount of data and the size of the moe?

Call the moe m. When planning our study we may want to specify the value of m. We do not want to change α[78], and σ is viewed as a property of nature and thus impossible to change. Thus we can only change n.

We can solve the equation for n and thus get a formula for the sample size needed to attain a margin of error m[79]:

[78] The relationship between α and m can be thought of as a trade-off between precision and accuracy. Here, *precision* refers to the size of the margin of error; it describes how well we know the location of the parameter of interest. We like being precise. We can gain precision by sacrificing *accuracy*, which is how likely the CI achieves its goal of containing the parameter of interest. While we want to be precise, we also want to be accurate, and wider intervals are naturally more accurate, all else being equal (or *ceteris paribus*, as the economists like to say). The only way to gain precision without sacrificing accuracy is increasing the sample size, n.

[79] The textbook has a similar formula but it involves the **width** of the CI, which is $w = 2m$. I prefer to use the margin of error here.

Example 1

An automated assembly line producing ball bearings should produce bearings with a diameter of 5mm. Quality control personnel run the line and get a sample of ten bearings. The bearings are known to have a standard deviation of $\sigma = 0.1$ mm[80]. The measured ball bearing diameters are listed below:

```
bearings <- c(10.396, 10.497, 10.655, 10.578, 10.543,
              10.575, 10.563, 10.549, 10.546, 10.489)
mean(bearings)
```

```
## [1] 10.5391
```

1. Construct a 95% CI for the mean diameter of the ball bearings.

2. Management is not satisfied with the margin of error, and want an estimate accurate up to 0.01 mm. Find a sample size n that attains this (while using a 95% CI).

[80] This assumption is clearly unrealistic; not only that, the mean μ is usually known before σ is, as you should expect from your study of probability and the nature of σ; thus its unlikely to see a study where σ is known but not μ. We will see in the next section what happens when we drop this assumption, but if n is large, you could replace σ with the sample standard deviation s and still get a quality CI, thanks to the law of large numbers and a result known as Slutzky's Theorem (Slutsky, 1925).

```r
suppressPackageStartupMessages(library(BSDA))
z.test(bearings, sigma.x = 0.1)
```

```
##
##  One-sample z-Test
##
## data:  bearings
## z = 333.28, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  10.47712 10.60108
## sample estimates:
## mean of x
##   10.5391
```

```r
ceiling((qnorm(1 - .05/2) * 0.1 / 0.01)^2)  # Needed n
```

```
## [1] 385
```

In many cases we can get formulas for confidence intervals that are either exact (if the assumptions hold) or approximately accurate for large n. This is not always the case, though, and we may need to use numerical techniques, such as **bootstrapping**, to get confidence intervals. This involves resampling the data and computing an estimate of the parameter of interest, $\hat{\theta}$, many times to get an estimate of the sampling distribution of $\hat{\theta}$. The percentiles of the simulated data can then be used to form the confidence interval.

Example 2

1. Use bootstrapping to estimate a 95% CI for the mean ball bearing diameter mentioned in Example 1.

```r
(xbar <- mean(bearings))  # Estimate
```

```
## [1] 10.5391
```

```r
xstars <- replicate(1000, {  # Simulations
  sim_bearings <- sample(bearings, size = 10, replace = TRUE)
  mean(sim_bearings)
})
head(xstars)
```

```
## [1] 10.5606 10.5395 10.5509 10.5280 10.5219
## [6] 10.5170
```

```
(xbarstar <- mean(xstars))  # Mean of simulated means
```

```
## [1] 10.53926
```

```
## Percentiles of simulated means
(xbar_perc <- quantile(xstars - xbarstar, c(0.025, 0.975)))
```

```
##       2.5%      97.5%
## -0.0396615  0.0379435
```

```
(xbar + xbar_perc)  # Bootstrap-estimated CI
```

```
##     2.5%    97.5%
## 10.49944 10.57704
```

2. Repeat the above procedure for the sample median. Which interval is more precise?

```
## Below I committed a programming sin: copy/paste programming!
## I should have written a function to generalize the
## procedure. But I have other goals, such as showing the
## intermediate steps.
(xtilde <- median(bearings))  # Estimate
```

```
## [1] 10.5475
```

```
xstars2 <- replicate(1000, {  # Simulations
  sim_bearings <- sample(bearings, size = 10, replace = TRUE)
  median(sim_bearings)
})
head(xstars)
```

```
## [1] 10.5606 10.5395 10.5509 10.5280 10.5219
## [6] 10.5170
```

```
(xtildestar <- mean(xstars2))  # Mean of simulated medians
```

```
## [1] 10.5469
```

```
## Percentiles of simulated medians
(xtilde_perc <- quantile(xstars2 - xtildestar, c(0.025, 0.975)))
```

```
##      2.5%     97.5%
## -0.053901  0.028099
```

```
(xtilde + xtilde_perc)  # Bootstrap-estimated CI
```

```
##    2.5%   97.5%
## 10.4936 10.5756
```

```
## Compare widths
(w1 <- diff(xbar_perc))  # Ignore the column name; not informative here
```

```
##    97.5%
## 0.077605
```

```
(w2 <- diff(xtilde_perc))  # Wider
```

```
## 97.5%
## 0.082
```

```
(w2 / w1 - 1) * 100  # The percentage by which the second interval is larger
```

```
##    97.5%
## 5.663295
```

Section 2: Large-Sample Confidence Intervals for a Population Mean and Proportion

The assumption that we know σ is clearly unrealistic. If n is large, though[81], we can replace σ[82] with the sample standard deviation, s. This is because of the following:

Proposition 15. *For a collection of i.i.d.r.v. $X_1, \ldots, X_n$ with sample mean $\bar{X}$ and sample standard deviation S, if $\mathbb{E}[X_1] = \mu$ and $\text{Var}(X_1) < \infty$, for n large, the approximate distribution of $Z = \frac{\bar{X}-\mu}{S/\sqrt{n}}$ is the standard Normal distribution.*

Thus we have the (approximate) $100(1-\alpha)\%$ CI:[83]

[81] As a rule of thumb, we can consider $n > 40$ as "large" in this context.

[82] In this context statisticians view σ as a **nuisance parameter**. We are not interested in the value of σ, but in order to make a statement about μ we are forced to estimate it.

[83] The quantity $s/\sqrt{n}$ is called the **standard error** of the mean, since it estimates the mean's standard deviation.

How would we go about sample size planning in this case? Our formulas seem to require future information. The easiest approach is to guess σ, erring on the side of large values as large σ yield larger n and thus smaller margin of errors.[84]

[84] This ethos of this approach is known as being "conservative", since we are trying to err on the side of more precision than desired. In this case, we err on the side of collecting more data than needed rather than collect too little and get a margin of error that is larger than desired.

Example 3

At the behest of management a new sample of ball bearings was collected, this time with $n = 61$ (people decided that 385 ball bearings were too many; the study should not cost that much money). The new sample mean is $\bar{x} = 10.488$ mm, and the sample standard deviation is $s = 0.105$ mm. Compute a 95% confidence interval for the mean diameter μ. Based on this CI, is it plausible the assembly line does not produce ball bearings of the desired diameter of 10 mm?

```r
xbar <- 10.488
s <- 0.105
n <- 61
(m <- qnorm(0.975) * s / sqrt(n))  # moe
```

```
## [1] 0.02634951
```

```r
c(xbar - m, xbar + m)
```

```
## [1] 10.46165 10.51435
```

Confidence intervals have a close cousin, called **confidence bounds**[85]. The number $l(x_1, \ldots, x_n)$ is a $100(1-\alpha)$**% confidence lower bound** for a parameter θ if

$$\mathbb{P}\left(l(X_1, \ldots, X_n) \leq \theta\right) = 1 - \alpha$$

Similarly, $u(x_1, \ldots, x_n)$ is a $100(1-\alpha)$**% confidence upper bound** for a parameter θ if

$$\mathbb{P}\left(\theta \leq u(X_1, \ldots, X_n)\right) = 1 - \alpha$$

We have the following large-sample confidence bounds for the population mean μ

[85] Confidence bounds can be viewed as confidence intervals with one of the end points being infinite.

Example 4

The stock with ticker symbol CGM had an average daily return of 0.07% over the last 200 days, with a standard deviation of 0.8%. Compute a 99% confidence lower bound for the mean return of the stock.

```
0.07 - 0.8 * qnorm(.99)/sqrt(200)
```

```
## [1] -0.06159811
```

Up until now we have been working with continuous data and our objective was to describe the location of the mean μ of the data. Suppose instead that we are working with binary/Bernoulli data and want to estimate the population proportion p of "successes". We can find a confidence interval for p[86] by working with

[86] The problem of finding a confidence interval for p demonstrates how many different procedures can be used to get different results intended to solve the same problem. Wikipedia (2018) lists eight different intervals CIs for p. The traditional CI used was

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

but this interval exhibits pathological behavior for strange combinations of n and p. The interval recommended in this class is known as the Wilson score interval, which biases the parameter estimate slightly to 0.5. An interval not mentioned in Wikipedia (2018) is the CI obtained when adding two "imaginary" successes and two "imaginary" failures to the sample; this interval seems to work well.

After isolating p in the inequality so that it's bounded by two computable numbers requiring only a sample of data, we get the following confidence interval:

We can turn the CI into a confidence bound by replacing $\alpha/2$ with α and $\pm$ with $+$ or $-$, depending on whether we want an upper bound or lower bound.

Prior to our study, if we want to choose a sample size n to achieve a moe m, our sample size should be

Here, $\tilde{p}$ is a *guess* at what the population proportion will be. If we are uncomfortable with making a guess, use $\tilde{p} = 0.5$; this will maximize m and guarantee that the observed moe will not exceed m (this is the most conservative approach). If we have a belief about the location of p we could economize during data collection somewhat by choosing $\tilde{p}$ to be near our belief, bearing in mind that the close $\tilde{p}$ is to 0.5, the larger our sample size (and smaller our observed moe) will be.

Example 5

Jack Johnson and John Jackson are running for mayor of New New York. The Johnson campaign conducts a survey of voters to deter-

mine who they support in the upcoming election.

1. The Johnson campaign will be constructing a 95% CI and does not want the moe to exceed 0.03 (or 3%). What sample size does the campaign need to achieve this?

2. In aa sample of 1068 New New York voters, 560 reported they planned to vote for Jack Johnson. Construct a 95% CI for the proportion of voters supporting Johnson. Based on the CI, who is winning?

```r
suppressPackageStartupMessages(library(Hmisc))
ceiling((qnorm(.975) * 0.5 / 0.03)^2)  # Sample size
```

```
## [1] 1068
```

```r
binconf(560, 1068, alpha = 0.05, method = "wilson")  # CI
```

```
##   PointEst     Lower     Upper
##  0.5243446 0.4943595 0.5541551
```

Section 3: Intervals Based on a Normal Population Distribution

From this point on in the chapter, we will assume that our data is an i.i.d. random sample from a *Normal* distribution with unknown mean and standard deviation. The intervals mentioned in Section 2 work for any underlying distribution so long as n is large enough. Here, we want intervals for when n is not considered large. The procedures mentioned in this section often work fine when n is large and the data doesn't follow a Normal distribution, though.

We start with the following theorem:

Theorem 5. *Suppose $\bar{X}$ is the sample mean of n i.i.d. Normal random variables with mean μ and S is the sample standard deviation. The random variable*

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

follows a t distribution with $\nu = n - 1$ degrees of freedom (denoted $T \sim t(n-1)$).

The $t(\nu)$ distribution[87] is a probability distribution with the following properties:

[87] This distribution is often called Student's t distribution in honor of the pseudonym of William Gosset. Gosset was employed by Guinness (the brewer), and at the time Guinness was engaged in a program to make beer brewing scientific. Eventually the experiments Guinness's burgeoning R & D department wanted to conduct required statistical methods that did not yet exist, so Gosset, then one of their brewers, began studying statistics and mathematics to develop methods

The t distribution depends on a parameter known as the **degrees of freedom (df)**. This name comes from the fact that among the n deviations $X_1 - \bar{X}, \ldots, X_n - \bar{X}$, the condition $\sum_{i=1}^{n}(X_1 - \bar{X}) = 0$ means only $n-1$ of these deviations are freely determined.

The t **critical value** $t_{\alpha,\nu}$ satisfies

Table A.5 gives critical values for the t distribution for various α and ν.

Confidence intervals based on the t distribution resemble those from the previous section, but with z_α replaced with $t_{\alpha,n-1}$.

We can get confidence bounds rather than confidence intervals by replacing $\pm$ with either $+$ or $-$ and $t_{\alpha/2,n-1}$ with $t_{\alpha,n-1}$.

Since we assume the data follows a Normal distribution, we should check that this assumption is reasonable for our dataset. Techniques for checking the normality assumption range from probability plots to box plots to statistical tests. Use whatever method you prefer.

Example 6

Assume that the diameter of the ball bearings from Example 3 follow a Normal distribution. Compute the requested CI but using the t distribution. Compare to the CI found in Example 3.

```
(m2 <- qt(0.975, df = n - 1) * s / sqrt(n))  # moe
```

```
## [1] 0.02689175
```

```
c(xbar - m2, xbar + m2)
```

```
## [1] 10.46111 10.51489
```

There are other satistical intervals than confidence intervals. A **prediction interval (PI)** is an interval intended to describe the range of values that will likely include a future observation. If we denote our future observation with X_{n+1} the interval $(l(x_1, \ldots, x_n), u(x_1, \ldots, x_n))$ is a $100(1-\alpha)\%$ PI if

For Normally distributed data our PI is given below:

Again, we can get formulas for prediction upper bounds or prediction lower bounds with the usual substitutions.

Example 7

Over the past 121 days, the daily percentage change of the price of the stock with ticker symbol CGM had the following sample mean and standard deviation:

```
mean(cgm)
```

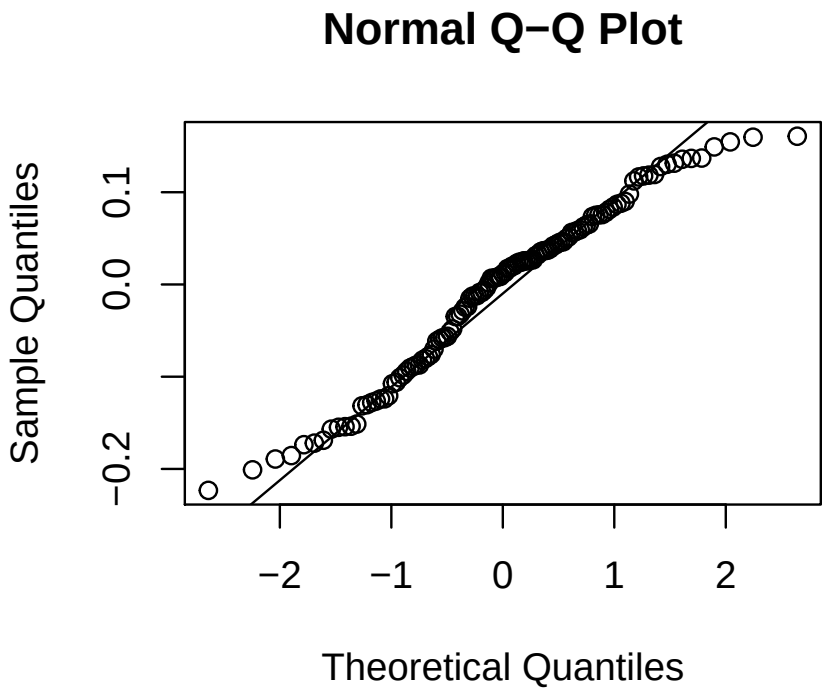
```
## [1] -0.005115001
```

```
sd(cgm)
```

```
## [1] 0.0922781
```

A look at these daily returns' probability plot suggests that we can reasonably assume that the price fluctuations follow a Normal distribution[88]:

```
qqnorm(cgm)
qqline(cgm)
```



Construct a 99% prediction lower bound for price movements.

[88] Actual asset price fluctuations are usually *not* Normally distributed. Instead, asset price fluctuations exhibit "heavy tails"; that is, extreme price movements are far more likely than the Normal distribution would suggest. Nevertheless, many models in finance for asset prices assume that price fluctuations follow a Normal distribution. See Mandelbrot and Hudson (2007) to learn more.

```
mean(cgm) - sd(cgm) * qt(.99, df = length(cgm) - 1 * sqrt(1 + 1/length(cgm)))
```

```
## [1] -0.2226907
```

Confidence intervals are meant to capture the mean and prediction intervals are meant to capture future values. **Tolerance intervals** are intervals such that at least k% of the population should be between the bounds of the interval; this statement is made with confidence level $100(1-\alpha)\%$[89].

[89] For example, we may have an interval such that, with 95% confidence, 99% of the population is within the bounds of the interval

The visualization of what is done by a tolerance interval is given below:

Tolerance intervals take the form

We still have the obvious translation to tolerance bounds. Tolerance critical values are given in Table A.6 in the textbook.

Example 8

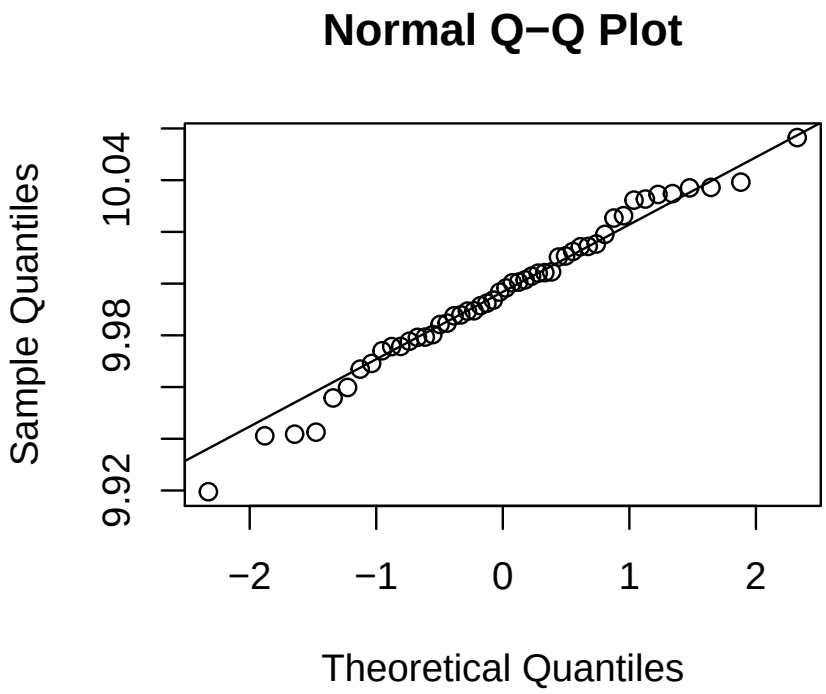
In light of previous studies, management has instructed the assembly line producing 10mm ball bearings to retool. After the retooling a sample of 50 ball bearings is produced by the line. Management will be satisfied if 99% of ball bearings produced by the line have a diameter that is within 0.1mm of the specified diameter of 10mm. Construct a 99% tolerance interval for the diameter of the ball bearings that is correct with 95% confidence, using the following data.

```
bearings2
```

```
##  [1] 10.001461 10.034805 10.014253  9.955770
##  [5] 10.012418  9.975701 10.000594 10.000315
##  [9] 10.010690 10.004044 10.015320 10.014393
```

```
## [13]  9.959830  9.987546  9.941776 10.026255
## [17] 10.025361 10.002873 10.019028 10.056476
## [21] 10.037196 10.004245 10.037012  9.974021
## [25] 10.039246  9.993619  9.996703  9.980263
## [29]  9.987879  9.942542  9.991442  9.941127
## [33]  9.975634  9.984178  9.989484 10.032692
## [37]  9.979320  9.977702 10.010261 10.034477
## [41] 10.004526  9.998308  9.992430  9.979205
## [45]  9.966931  9.919507  9.969121  9.989352
## [49] 10.032301  9.984713
```

```
qqnorm(bearings2)
qqline(bearings2)
```




```
mean(bearings2)
```

```
## [1] 9.996087
```

```
sd(bearings2)
```

```
## [1] 0.02894869
```

```
## For constructing tolerance intervals
suppressPackageStartupMessages(library(tolerance))
normtol.int(bearings2, alpha = .05, P = 0.99, side = 2)

##   alpha    P    x.bar 2-sided.lower
## 1  0.05 0.99 9.996087      9.905473
##   2-sided.upper
## 1       10.0867
```

What do you do when you don't have Normally distributed data and n is not large? This depends on what you are attempting to do. Some procedures, such as the t procedures for constructing confidence intervals, are **robust** to non-normality in some contexts; that is, failure of holding to the assumption does not seem to change the end result very much. But prediction intervals and tolerance intervals are *not* robust to the normality assumption and you may need to an interval constructed for a more appropriate distribution. Bootstrapping and other non-parametric procedures (not discussed in this course) could also provide a solution. Perhaps consider reading the book Hahn and Meeker (2011) to learn about other intervals that may be useful for your problem.

Section 4: Confidence Intervals for the Variance and Standard Deviation of a Normal Population

We may be interested in constructing a confidence interval for the population variance σ^2 or standard deviation σ. We will be keeping the assumptions made in Section 3; in fact, those assumptions are more crucial. Not only are the procedures I will suggest *not* robust to the Normality assumption, if our data isn't Normally distributed, we may not even consider σ a good measure of spread in the data (especially if our underlying distribution is not symmetric).

Theorem 6. *Suppose $\bar{X}$ is the sample mean of n i.i.d. Normal random variables with mean μ and S^2 is the sample variance The random variable*

$$\frac{(n-1)S^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

follows a $\chi^2(n-1)$ distribution.

Let $\chi^2_{\alpha,\nu}$ satisfy

We can derive the CI for σ^2 by working with

The resulting CI is given below:[90]

[90] Notice this is *not* an equal-tail interval!

We can get a CI for σ by taking the square root of the lower and upper bounds. We can get one-sided intervals by using either the upper or lower bound exclusively and replacing $\alpha/2$ with α.

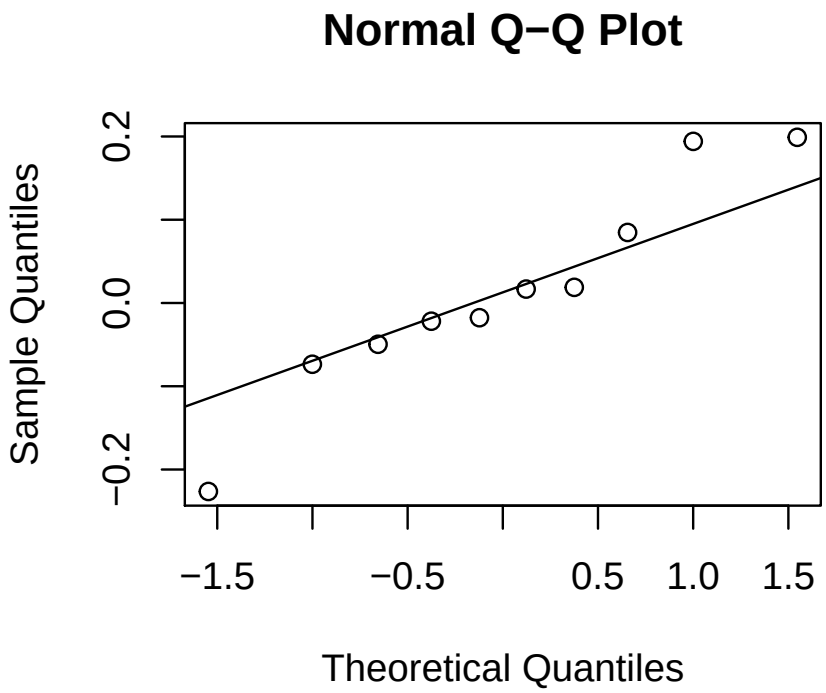
Example 9

We have the following returns from the previous ten days of the stock with ticker symbol CGM:

```
cgm2 <- c(-0.2264,  0.0188, -0.0496,  0.1990,  0.1941,
          -0.0219, -0.0177,  0.0847,  0.0167, -0.0736)
```

Based on the plot below the returns seem to follow a Normal distribution:

```
qqnorm(cgm2)
qqline(cgm2)
```



The standard deviation and variance of the stock's daily returns are given below:

```
var(cgm2)
```

```
## [1] 0.01594101
```

```
(vol <- sd(cgm2))
```

```
## [1] 0.1262577
```

Construct a 90% CI for the true σ[91] of the stock's returns.

[91] In finance, σ is frequently referred to as the *volatility* of the asset's price.

```
n <- length(cgm2)
(l <- (n - 1) * vol^2 / qchisq(.05, df = n - 1,
                               lower.tail = FALSE))  # Variance lower bound
```

```
## [1] 0.008479775
```

```
(u <- (n - 1) * vol^2 / qchisq(1 - .05, df = n - 1,
                               lower.tail = FALSE))  # Upper bound
```

```
## [1] 0.04314715
```

```
c(sqrt(l), sqrt(u))  # Bounds for the standard deviation
```

```
## [1] 0.0920857 0.2077189
```

Chapter 8: Tests of Hypotheses Based on a Single Sample

Introduction

Statistics involes more than parameter estimation. We may not care about the actual value of the parameter but rather whether the parameter is a particular value or within some range. In this case we may prefer to perform a statistical hypothesis test rather than construct a confidence interval.

In this chapter we see for the first time statistical hypothesis testing, involving only a single sample. The hypotheses of interest will typically be making a statement about the value of a parameter, though other hypothesis tests make more general statements. The fundamental principles, though, are the same, along with the general format of a test.

Hypothesis testing is a popular procedure, which suggests it's also frequently abused. We should always remember that hypothesis testing is part of our toolset for reaching conclusions about a phenomenon using a dataset; it is not the *only* tool that should be used We should supplement hypothesis testing with other procedures, such as visualization and providing point estimates. Furthermore, we should be *honest* when collecting our data and be sure we are not "coercing" the dataset to get an answer we want.[92]

[92] As Nobel Prize winning economist Robert Coase said, "If you torture the data enough, nature will always confess."

Section 1: Hypotheses and Test Procedures

A **statistical hypothesis** is a statement about the probabilistic properties of a data-generating process.[93] A **test of hypotheses** is a procedure where sample data is used to decide which of two competing hypotheses better describes the process that generated the data. The **null hypothesis** (usually denoted H_0) can be thought of as the current assumption about the data[94], while the **alternative hypothesis** (usually denoted H_A) is the assumption that will replace the null hy-

[93] This is usually a statement about a parameter, a collection of parameters, or even whether the data follows some distribution.

[94] Usually H_0 is the statement we seek to disprove, but this is not always the case; for example, tests for distribution, which intend to determine if the data follows a particular distribution, will often state that under the null hypothesis the data follows the distribution of interest.

pothesis if we reject the null hypothesis. If we don't reject H_0, we do not say that we accept H_0 but rather that we failed to reject H_0.

Hypothesis testing is a form of *reductio ad absurdum* ("argument to absurdity"), similar to a proof by contradiction; the argument is that by assuming the null hypothesis is true we see a result in the data that is "absurd", so we should surrender our belief in H_0. If this "absurdity" in the data does not appear, though, that does not mean H_0 is true; it just means we could not show it is false, or that H_A is more correct.[95]

In this chapter we consider tests that make statements about a population parameter θ. These tests almost always take the following form in practice:

We call θ_0 the **null value** for θ, and it is the assumed value of θ under H_0.[96]

Statistical tests (of any form) follow the procedure described below:

1. Identify H_0 and H_A.
2. Specify a number $\alpha \in (0,1)$, usually small (typical α are $\alpha \in \{0.1, 0.05, 0.01, 0.001\}$; there is an interpretation of α I will explain later that can guide this decision). This is called the **significance level** of the test.
3. Collect data and compute the test statistic; call the random version of the test statistic T for now, and let the observed (computed) value of T be $\hat{T}$. If H_0 is true, the distribution of T is known.
4. Compute a quantity known at the **p-value**, denoted here p_{val}[97]. The definition of p_{val} in general is[98]

$$p_{\text{val}} = \mathbb{P}\left(\text{Observe } T \text{ more contradictory to } H_0 \text{ than } \hat{T}\right)$$

5. If $p_{\text{val}} < \alpha$, reject H_0; otherwise ($p_{\text{val}} \geq \alpha$) do not reject H_0. (Because of this rule, p_{val} is sometimes referred to as the **observed significance level** of the test, as it is the smallest α at which you would reject H_0.)
6. Conclude and interpret the results of the test.

[95] The usual comparison is the ancient Roman legal principle (still in use in America) of "innocent until proven guilty"; we assume that the individual on trial is innocent (i.e. H_0 is "true"), and the burden of proof lies on the prosecution (the statistical test) to show that this assumption is "absurd" based on the evidence (the data) and we should then assume the individual is guilty (H_0 is "false", and H_A better describes reality). Failure to prove guilt, though, does not imply innocence, and the "beyond reasonable doubt" criteria sets a high bar for proving guilt. It tilts justice in favor of letting guilty people go (a Type II error) as opposed to using the state's resources to punish the innocent (a Type I error).

[96] We amost never see H_0 of the form

$$H_0 : \theta \leq \theta_0$$

or

$$H_0 : \theta \geq \theta_0$$

This is because the statistical test does not change if we replace the inequality with equality. The border value θ_0 is the most difficult case to check, and it can be shown that if we reject H_0 at the border we can safely reject for all other potential values of θ, while if we could not reject H_0 when assuming $\theta = \theta_0$ we should reject H_0 at all. Consequently we can view H_0 as actually making a statement about all possible θ within a region when H_A is one-sided.

[97] The usual notation for p-values is simply p, but we will run into situations in this class where the letter p appears in many places, so I use this notation to keep all these different p's straight.

[98] The classical approach to statistical testing does not involve p-values but instead a critical value, T_0, and if $\hat{T} > T_0$, H_0 would be rejected. This theory still underlies statistics; power/Type II error analysis and the formulas for computing p-values are derived with this theory in mind. However, there are advantages to referring to p-values. One is that software usually computes p-values. Another is that p-values have a universal interpretation; given any p-value you can decide whether to reject H_0 or not even if you don't know the context of the test. Additionally, readers can decide whether a reported p-value is convincing for themselves personally, regardless of what the authors of the

Be clear that p_{val} is the probability of observing a test statistic at least as contradictory to H_0 as the observed test statistic. If we were to say that large T are evidence against H_0 (with larger T meaning even more evidence against H_0), then $p_{\text{val}} = \mathbb{P}\left(T > \hat{T}\right)$; that is, it is the probability of seeing even more contradictory evidence than what was seen.

The following are *incorrect* interpretations of p_{val}:

- The probability H_0 is true or false.
- The probability the conclusion of the test is due to random chance alone.

Additionally, practitioners should not fret over exactly what threshold a p-value passes (such as whether $p_{\text{val}} < 0.05$). While (5) in the above description of the statistical testing procedure suggests that certain p-value imply certain conclusions, p-values are more useful when considered as a measure of how strong the evidence against H_0 is.[99]

[99] People have identified as one of the culprits of the so-called reproducibility crisis (many results in published scientific papers cannot be reproduced), and they are frequently abused. The problem has gotten severe enough that in 2006 the American Statistical Association (ASA) issued a statement about the appropriate use and interpretation of p-values (Wasserstein and Lazar, 2016). However, the problems associated with p-values can be pinned (more fairly) on publishing practices and how publication decisions are made. Journals are biased to "positive results" (i.e. when H_0 is rejected) and have given $\alpha = 0.05$ unreasonable importance. This can lead to malicious practices such as p-hacking (rephrasing a statistical problem until "statistically significant" results are found), or ignoring the size of the effect found in the paper. See Aschwanden (2015) and Aschwanden (2016) for interesting discussions and even interactive demonstrations of these issues.

In hypothesis testing, there are two types of errors. A **Type I error** is rejecting H_0 when H_0 is true, while a **Type II error** is failing to reject H_0 when H_0 is false. The table below visualizes the relationship:

Immediately after a test, you do not know whether you committed an error or what the nature of the error is. Error analysis is part of study design, conducted before any data is collected. It determines what must be observed to reject H_0 and what sample size the study should use. There should be a discussion about what happens when a Type I or Type II error is made, what the consequences are, the relative severity of the consequences, and thus what the acceptable error rates should be.

α is the Type I error rate:[100]

[100] Actually this is the case when the test statistic is a continuous variable. For discrete variables, we may choose a desired α but due to the discrete nature of the cdf the actual Type I error rate may be less than specified (when being conservative). We see this in Example 1.

In this context, the Type II error rate depends on what the true value of θ is; we call $\beta(\theta_A)$ the Type II error rate when $\theta = \theta_A$:

A related concept to the Type II error rate is the **power** of the statistical test; the power is the probability of rejecting H_0 when $\theta = \theta_A$. It is defined below:

Power relates to α and β in the following way:

There is a general relationship between α and β:

After we have reflected on the consequences of Type I and Type II errors, we may decide on an acceptable Type I error rate, α. We then focus on a particular θ_A we want our test to be able to detect and what the acceptable Type II error rate $\beta(\theta_A)$ should be. Once we have these pieces of information, we may find a sample size n that achieves these two error rates for our test.

Example 1

I claim that I am an 80% freethrow shooter, but you don't believe me; you think I make less than 80% of freethrows. To settle the dispute, we agree that I will shoot 20 free-throws and you will count how many baskets I manage to make. Based on this you will decide whether you believe my claim. You decide to use $\alpha = 0.05$ as your significance level.

1. Identify H_0 and H_A.

2. What is the test statistic? What is its distribution under H_0?

3. Out of 20 baskets, I manage to make 11. Compute p_{val}.

4. What is the conclusion of the test?

5. Let N_α denote the fewest number of baskets I could make while still allowing you to believe my claim when you use significance level α (that is, if $X \sim \text{BIN}(20, 0.8)$, N_α is the largest number such that $\mathbb{P}(X < N_\alpha) \leq \alpha$). Find $N_{0.05}$.[101]

[101] (5) and on are questions we would ask before we observed any data and reached a conclusion.

6. While $\alpha = 0.05$ is the specified Type I error rate, due to the discrete nature of the test statistic, it is not the actual Type I error rate. What is the actual Type I error rate?

7. Suppose I were not an 80% freethrow shooter and instead only make 75% of my baskets. What is the Type II error rate in this case?

```r
pbinom(11, 20, .8)  # Part 3
```

```
## [1] 0.009981786
```

```r
(N <- qbinom(.05, 20, .8) - 1)  # Part 5
```

```
## [1] 12
```

```r
pbinom(N, 20, .8)  # Part 6
```

```
## [1] 0.03214266
```

```r
pbinom(N, 20, .75, lower.tail = FALSE)
```

```
## [1] 0.8981881
```

Example 2

Let μ denote the population mean. We wish to determine if the true population mean is greater than the specified value μ_0.

1. State the null and alternative hypothesis.

2. We collect a dataset $X_1, \ldots, X_n$ from the population, with $\mathbb{E}[X] = \mu$, and $\text{SD}(X) = \sigma$. Consider the test statistic

$$Z = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}}$$

According to the central limit theorem, what is the approximate distribution of Z under H_0?

3. What is the (approximate) distribution when H_A is true and $\mu = \mu_A > \mu_0$?

4. Let α be the level of significance of the test. We will reject H_0 when Z is larger than some threshold value. Find this threshold value such that the Type I error rate is α.

5. Compute $\beta(\mu_A)$. This is the probability of *not* rejecting H_0 when $\mu = \mu_A$; in other words, Z is less than the threshold value even though $\mu = \mu_A > \mu_0$.

6. Given the answers to (4) and (5), find a sample size n such that a test with Type I error rate α will have Type II error rate β when $\mu = \mu_A$.

7. Let's now suppose that we are investigating whether men's average height is 5.9 ft., and under the alternative hypothesis men are taller than 5.9 ft. Phrase H_0 and H_A below.

8. Let our significance level be $\alpha = 0.1$. The standard deviation of height is known to be $\sigma = 0.5$. Suppose the true mean height for men is 6 ft. What is the Type II error rate when $n = 100$? Repeat for a potential mean height of 6.5 ft.

9. Find the sample size that, for a test with $\alpha = 0.1$, would have a Type II error rate of $\beta = 0.1$ when the true average height is 6 ft.

10. A sample mean height of 5.97 ft. is observed, and the sample size is the one found in part (9) above. Compute p_{val}.

11. Based on this data, what is the conclusion of the test?

```r
alpha <- 0.1
beta <- 0.1
sigma <- 0.5
mu0 <- 5.9
muA <- 6.0
xbar <- 5.97
(za <- qnorm(alpha, lower.tail = FALSE))
```

```
## [1] 1.281552
```

```r
(zb <- qnorm(beta, lower.tail = FALSE))
```

```
## [1] 1.281552
```

```r
pnorm(za + (mu0 - muA)/(sigma/sqrt(100)))  # Part 8
```

```
## [1] 0.2362404
```

```r
pnorm(za + (mu0 - 6.5)/(sigma/sqrt(100)))
```

```
## [1] 4.169523e-27
```

```r
(n <- ceiling((sigma * (za + zb) / (mu0 - muA))^2))  # Part 9
```

```
## [1] 165
```

```r
(z <- (xbar - mu0)/(sigma/sqrt(n)))  # Part 10
```

```
## [1] 1.798333
```

```r
(pval <- pnorm(z, lower.tail = FALSE))
```

```
## [1] 0.03606216
```

```r
(pval < alpha)  # Part 11
```

```
## [1] TRUE
```

Section 2: z Tests for Hypotheses about a Population Mean

From here, in order to perform a hypothesis test, we only need the following bits of information:

- The null hypothesis H_0, and potential H_A
- Assumptions about the data made by the test
- The test statistic and how to compute it
- How to compute p_{val} based on the test statistic

Our first case is a test for the mean μ when σ is known. This test is exact when the data was drawn from a Normal distribution, and asymptotically correct when the data is not Normally distributed.

Suppose σ is not known. If n is large[102], we can replace σ with the sample standard deviation S and thus use the test statistic

[102] Let's say $n > 40$.

The test is otherwise the same.

Below are formulas for computing Type II errors. If σ is not known, you will need to guess it.

The formulas below allow for sample size planning. Overestimating σ will produce large n and thus produce tests that may do better than specified.

Example 3

A factory that produces ball bearings is testing its assembly line to see whether the line produces ball bearings with the specified diameter of 10 mm or whether the line is not properly calibrated. The managers believe that the standard deviation of bearings produced by this line is $\sigma = 0.1$ mm. They want tests that are significant at the $\alpha = 0.01$ significance level.

1. State the null and alternative hypothesis.

2. What is the probability of a Type I error?

3. What is the probability of a Type II error when the ball bearings' mean diameter differ from the specified diameter by 0.1 mm and the sample size is $n = 20$?

4. The managers want to be able to detect a difference of 0.1 mm from the specified diameter with probability 0.9995. Find a sample size that guarantees this under our assumptions.

5. Using the sample size $n = 41$, experimenters run the line and produce some ball bearings. The following sample was observed:

```
bearings <- c(10.11, 9.858, 10.072, 10.007, 10.158, 9.878, 9.935, 9.787,
              9.993, 10.008, 9.927, 9.959, 10.086, 10.001, 9.881, 10.057, 9.913,
              9.744, 10.136, 10, 9.988, 10.022, 10.112, 10.013, 9.809, 10.014,
              10.036, 9.977, 9.952, 9.963, 9.955, 9.926, 10.095, 10.076, 9.994,
              9.93, 10.057, 9.923, 9.954, 9.969, 10.124)
mean(bearings)
```

```
## [1] 9.985341
```

```
sd(bearings)
```

```
## [1] 0.09381434
```

Using the sample standard deviation rather than σ, perform the appropriate statistical test to decide between H_0 and H_A, computing p_{val}.

6. Conclude.

```r
suppressPackageStartupMessages(library(BSDA))
alpha <- 0.01
beta <- 0.0005
sigma <- 0.1
mu0 <- 10
muA <- 10.1  # We could also use 9.9 and be fine
(za2 <- qnorm(alpha/2, lower.tail = FALSE))
```

```
## [1] 2.575829
```

```r
(zb <- qnorm(beta, lower.tail = FALSE))
```

```
## [1] 3.290527
```

```r
pnorm(za2 + (mu0 - muA)/(sigma/sqrt(10))) +
  pnorm(-za2 + (mu0 - muA)/(sigma/sqrt(10)))  # Part 3
```

```
## [1] 0.2787871
```

```r
(n <- ceiling((sigma * (za2 + zb)/(mu0 - muA))^2))  # Part 4
```

```
## [1] 35
```

```r
z.test(bearings, mu = 10, sigma.x = sd(bearings))  # Part 5
```

```
##
##  One-sample z-Test
##
## data:  bearings
## z = -1.0005, p-value = 0.3171
## alternative hypothesis: true mean is not equal to 10
## 95 percent confidence interval:
##   9.956625 10.014058
## sample estimates:
## mean of x
##  9.985341
```

Section 3: The One-Sample t Test

If we assume our data follows a Normal distribution, then the distribution of

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$$

is $t(n-1)$ when H_0 is true. Based on this we can describe a test based on the t distribution.

This test works better than the test described in the previous section when the data follows a Normal distribution, and the difference is noticeable for small n.[103]

Table A.5 isn't well suited for hypothesis testing; instead, use Table A.8.

Example 4

Repeat the test performed in Example 3 but using the t-test instead. Does your conclusion change?

[103] What's the difference between these two tests? What is the penalty for using the z-test rather than the t-test for Normally distributed data? Notice that $z_\alpha < t_{\alpha,n-1}$ for all n. Since the random variable T follows the $t(n-1)$ distribution, we can conclude that when we use the z-test instead of the t-test, p_{val} will be inappropriately small, and thus we are more likely to reject the null hypothesis. The true Type I error rate is *greater* than α! This phenomenon is known as **size inflation**. When n is large the inflation is negligible, but for small n it could be a problem.

```
t.test(bearings, mu = 10)
```

```
##
##  One Sample t-test
##
## data:  bearings
## t = -1.0005, df = 40, p-value = 0.3231
## alternative hypothesis: true mean is not equal to 10
## 95 percent confidence interval:
##   9.95573 10.01495
## sample estimates:
## mean of x
##  9.985341
```

Type II error analysis (including sample size planning) is more complicated for t-testing, and we do not have clean formulas like we did when σ was known. We either need to use software or graphs like those provided in Table A.17. When using software, the power $\pi(\mu_A)$ is usually referred to rather than $\beta(\mu_A)$, and the input is usually not μ_A but $d = (\mu_0 - \mu_A)/\sigma$. (Table A.17 also uses d.) Notice that a guess of σ needs to be made.

Example 5

Use Table A.17 to answer the following:

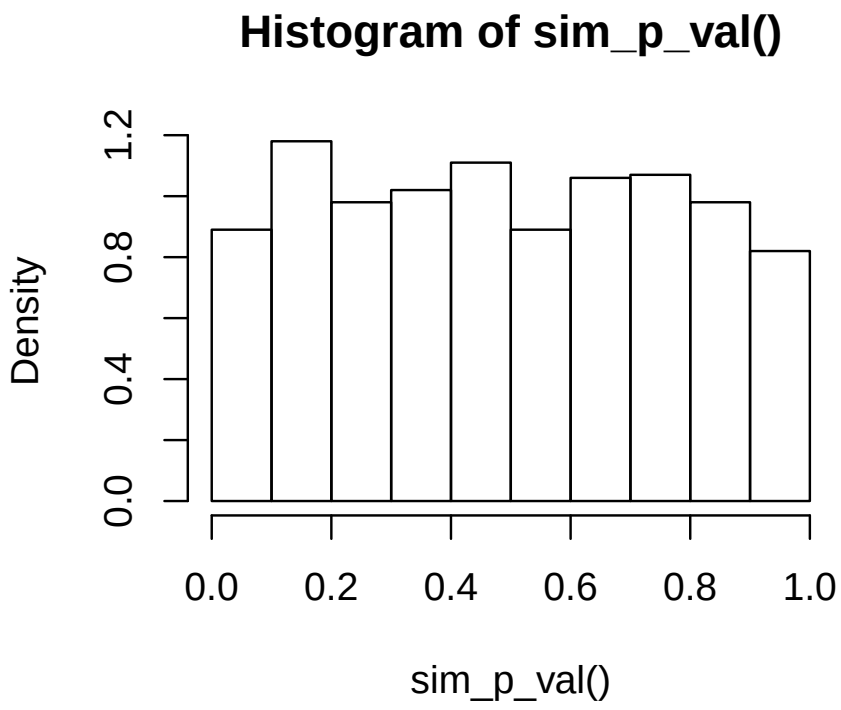
1. For a one-tailed t-test, what is the probability of a Type II error when the degrees of freedom is $\nu = 9$ and $|d| = 0.6$? Repeat with $\nu = 29$.

2. For a two-tailed test, what sample size is needed so that a test will have a Type II error rate of 0.1 when $|d| = 0.5$? Choose the smallest listed degrees of freedom.

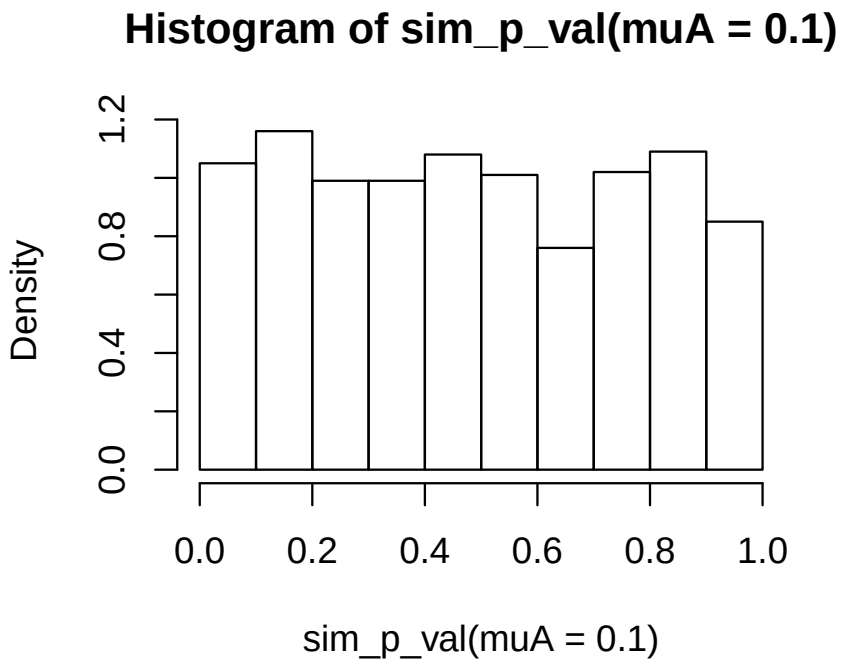
Something to consider when talking about p_{val}: this number is a statistic like any other quantity we compute from data, and thus it has a sampling distribution. Under the null hypothesis, if the assumptions of the t-test are met, then it can be shown that $p_{\text{val}} \sim \text{UNIF}(0,1)$. Under the alternative hypothesis, though, p_{val} follows a distribution other than $\text{UNIF}(0,1)$, and the sampling distribution concentrates near 0 as n grows or as Δ grows.

Below I simulate the distribution of p_{val} in different scenarios.

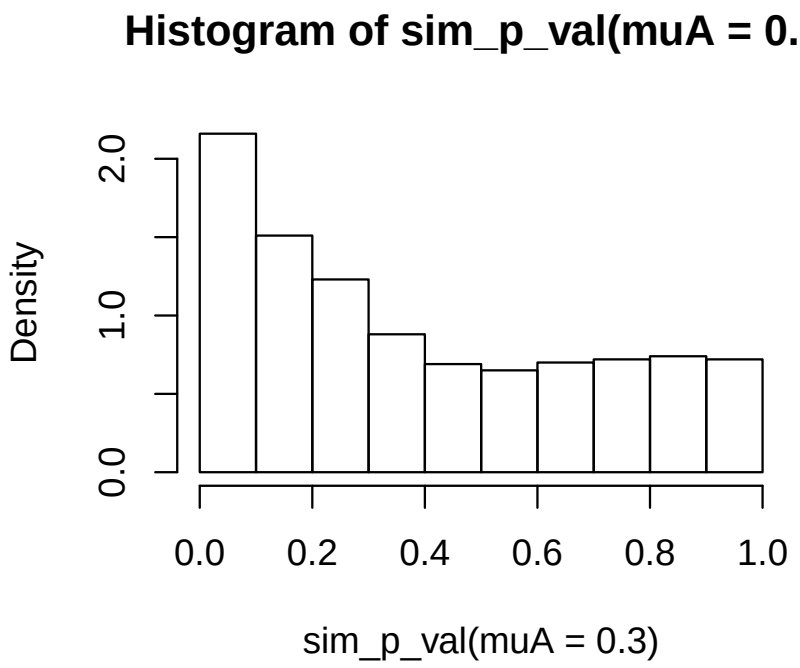
```
## I write a function to perform these simulations
sim_p_val <- function(M = 1000,    # Number of replications
                      mu0 = 0,     # Hypothesized mean
                      muA = NULL,  # True mean; if null, same as mu0
                      n = 10,
                      sd = 1,
                      alternative = c("two.sided", "less", "greater")) {
  if (is.null(muA)) {
    muA <- mu0
  }
  alternative <- alternative[1]

  replicate(M, {
    dat <- rnorm(n, mean = muA, sd = sd)
    return(t.test(dat, alternative = alternative, mu = mu0)$p.value)
  })
}

hist(sim_p_val(), freq = FALSE)
```



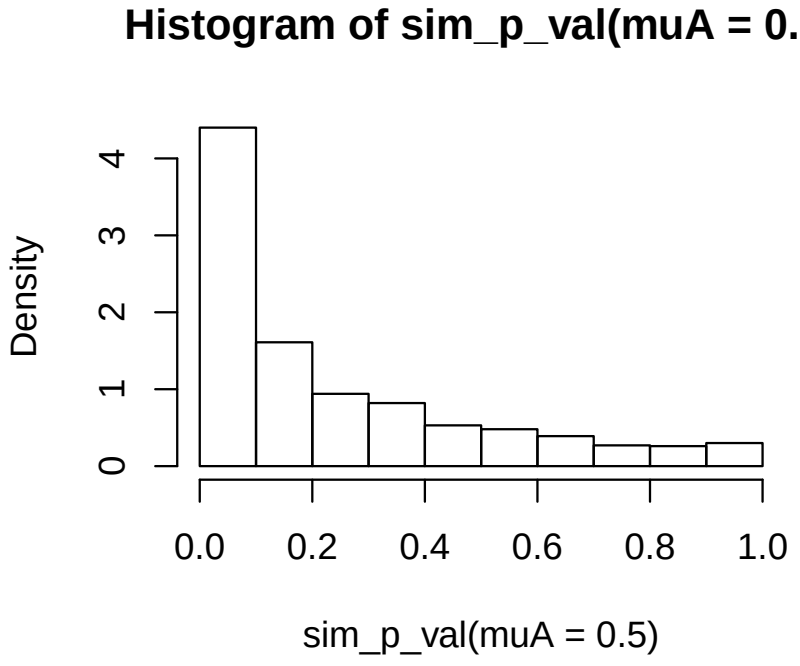

```
hist(sim_p_val(muA = 0.1), freq = FALSE)
```



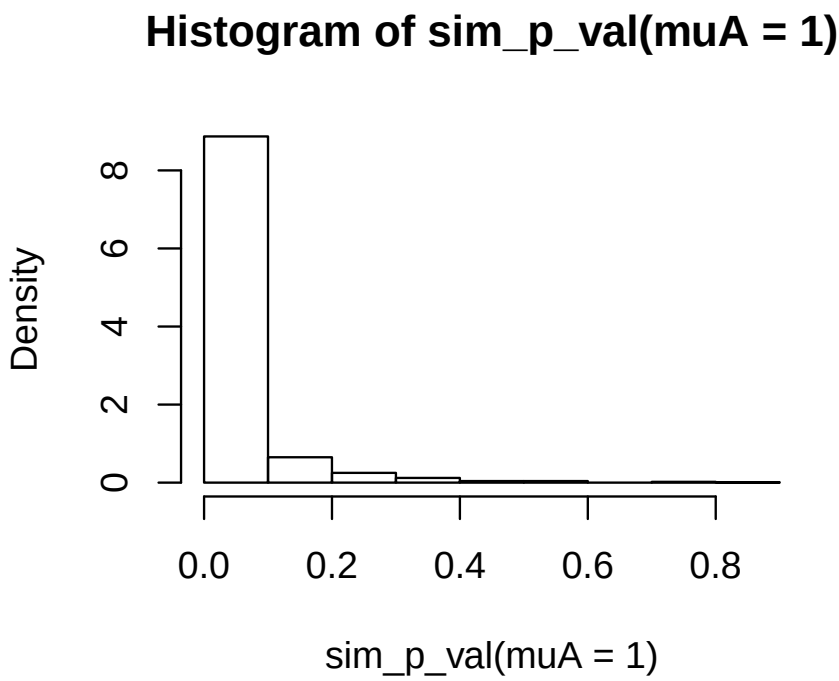

```
hist(sim_p_val(muA = 0.3), freq = FALSE)
```




```r
hist(sim_p_val(muA = 0.5), freq = FALSE)
```




```r
hist(sim_p_val(muA = 1), freq = FALSE)
```



When we compute a p_{val} and get a statistically significant result we may be interested in whether others repeating our study will also get a statistically significant result; in other words, whether they will be able to replicate our result. This issue is discussed in Boos and Stefanski (2011). They noted that for p-values that are near-misses (that is, $p_{\text{val}} < \alpha$ but only barely) there are good odds that replication studies will not also reject H_0, but when the p-value is much smaller than α, the odds of replication should be good. They even recommend reporting estimates of the replication probability to signal how fragile the results of the study are.

Section 4: Tests Concerning a Population Proportion

In Example 1 we saw what a small sample test for a population proportion looks like. When our data follows a Bernoulli distribution, we first state our null and alternative hypothesis:

Then we identify the distribution of the number of "successes" in the sample if H_0 is true:

Finally, we can provide a formula for computing p_{val}.

In this section I consider the large-sample version of the test. First, consider the sample proportion $\hat{p}$ computed from Bernoulli data $X_1, \ldots, X_n$, $X_i \sim \text{Ber}(p)$. Assume $H_0 : p = p_0$ is true. What then is $\mathbb{E}[\hat{p}]$ and $\text{Var}(\hat{p})$?

Based on this, what is the approximate distribution for $\hat{p}$ for large n?

Using this, we can create a large-sample test for sample proportions[104], described below.

[104] We can extend this reasoning to other statistics that asymptotically follow the Normal distribution. Suppose $\hat{\theta}$ is a consistent estimator of θ, and let $\text{SD}(\hat{\theta}) = \sigma_{\hat{\theta}}$. If we have

$$Z = \frac{\hat{\theta} - \theta}{\sigma_{\hat{\theta}}}$$

and the approximate distribution of Z is $N(0,1)$, then we can test $H_0 : \theta = \theta_0$ against some alternative using the statistic

$$Z = \frac{\hat{\theta} - \theta_0}{\sigma_{\hat{\theta}}}$$

In this case, under H_0, $\sigma_{\hat{p}} = \sqrt{np_0(1-p_0)}$, thus producing the large-sample test statistics described.

Below are large-sample Type II error analysis formulas:

Example 6

Jack Johnson and John Jackson are running for President of Earth. You work for the Johnson campaign and want to determine whether Johnson is currently the candidate with the most support. You plan on conducting a survey asking potential voters who they plan to vote for in the election.

1. Let p represent the proportion of potential voters who support Johnson. State an appropriate null and alternative hypothesis.

2. You want to be able to detect when Johnson has a 1% lead among potential voters with a probability of 95%. Find a sample size capable of achieving this when your level of significaince is $\alpha = 0.05$.

3. It turns out that Johnson is very rich and can afford to conduct a survey of the size found above, and in the survey of 61,000 potential voters, 30,698 reported they were planning on voting for Johnson. Compute p_{val} and conclude.

```
alpha <- 0.05
beta <- 0.05
p0 <- 0.5
pA <- 0.51
n <- 61000
x <- 30698
(za <- qnorm(alpha, lower.tail = FALSE))
```

```
## [1] 1.644854
```

```
(zb <- qnorm(beta, lower.tail = FALSE))
```

```
## [1] 1.644854
```

```
## Part 2
ceiling(((za * sqrt(p0 * (1 - p0)) + zb * sqrt(pA * (1 - pA)))/(pA - p0))^2)
```

```
## [1] 27051
```

```
prop.test(x, n, p = 0.5, alternative = "greater", correct = FALSE)
```

```
##
##  1-sample proportions test without
##  continuity correction
##
## data:  x out of n, null probability 0.5
## X-squared = 2.5708, df = 1, p-value =
## 0.05443
## alternative hypothesis: true p is greater than 0.5
## 95 percent confidence interval:
##  0.499916 1.000000
## sample estimates:
##         p
## 0.5032459
```

Section 5: Further Aspects of Hypothesis Testing

Suppose we want to perform hypothesis tests for describing the value of the population variance: that is, we wish to test

Assume that the data $X_1, \ldots, X_n$ is an i.i.d. sample from the $N(\mu, \sigma)$ distribution[105]. Then se can describe the distribution of S^2, the sample variance:

[105] The t-test we saw in Section 3 was somewhat robust to the Normality assumption, working well for large sample sizes even when the assumptions of the test are not met. However, the χ^2 test is *not* robust to this assumption. As discussed before, for non-Normal data, inference regarding σ^2 may not even be very useful when the data doesn't follow a Normal distribution.

Using this we can formulate a statistical test for inference for σ^2 (and thus σ as well):

Below are formulas for Type II error analysis:

Sample size planning for this test is difficult and may require the use of numerical techniques.

Example 7

1. Suppose we plan to use ten days of returns from a stock price series to test whether the volatility (that is, σ) of the stock is greater than 10% or not. State the null and alternative hypotheses.

2. Suppose the true volatility of the stock is 15%. Estimate the probability of comitting a Type II error when $n = 10$ and $\alpha = 0.1$.

3. We have the following returns from the previous ten days of the stock with ticker symbol CGM:

```
cgm2 <- c(-0.2264,  0.0188, -0.0496,  0.1990,  0.1941,
          -0.0219, -0.0177,  0.0847,  0.0167, -0.0736)
```

The standard deviation and variance of the stock's daily returns are given below:

```
var(cgm2)
```

```
## [1] 0.01594101
```

```
(vol <- sd(cgm2))
```

```
## [1] 0.1262577
```

Test whether the volatility (that is, σ) of the stock is greater than 10% or not with significance level $\alpha = 0.1$.

```
sigma0 <- .1
## Part 2
pchisq(sigma0^2/.15^2 * qchisq(.1, df = 10 - 1, lower.tail = FALSE),
       df = 10 - 1)
```

```
## [1] 0.3136713
```

```
## Part 3
pchisq((10 - 1) * var(cgm2)/sigma0^2, df = 10 - 1, lower.tail = FALSE)
```

```
## [1] 0.1105089
```

In hypothesis testing, we can find **statistically significant** results (where H_0 was rejected) that are not **practically significant**. That is, we might conclude that H_0 is false, but the difference between θ_0 and our best estimate of the true value of the parameter of interest are barely worth mentioning. Large sample sizes produce tests so powerful they can detect even tiny divergences from H_0, even if the actual effect is barely worth mentioning. Thus we should be cautious and not overstate the importance of our test's conclusions.

Example 8

Suppose we are testing to see if the proportion of individuals who have some rare disease is more than $p = 0.007$. We have a lot of funding and conduct a massive study and can conclude that, in fact, the true proportion of the population with the disease is more than 0.007. But our point estimate for this proportion is $\hat{p} = 0.00711$; this is barely larger than the hypothesized value, so the test's results are not noteworthy.

Statistical tests and confidence intervals have a connection. If we have a $100(1-\alpha)\%$ confidence interval $(l(x_1, \ldots, x_n), u(x_1, \ldots, x_n))$ and consider the set of hypotheses:

The CI can be interpreted as the set of θ_0 for which H_0 would not be rejected at significance level α. $100(1-\alpha)\%$ confidence bounds have a similar interpretation for the alternative hypotheses:

Example 9

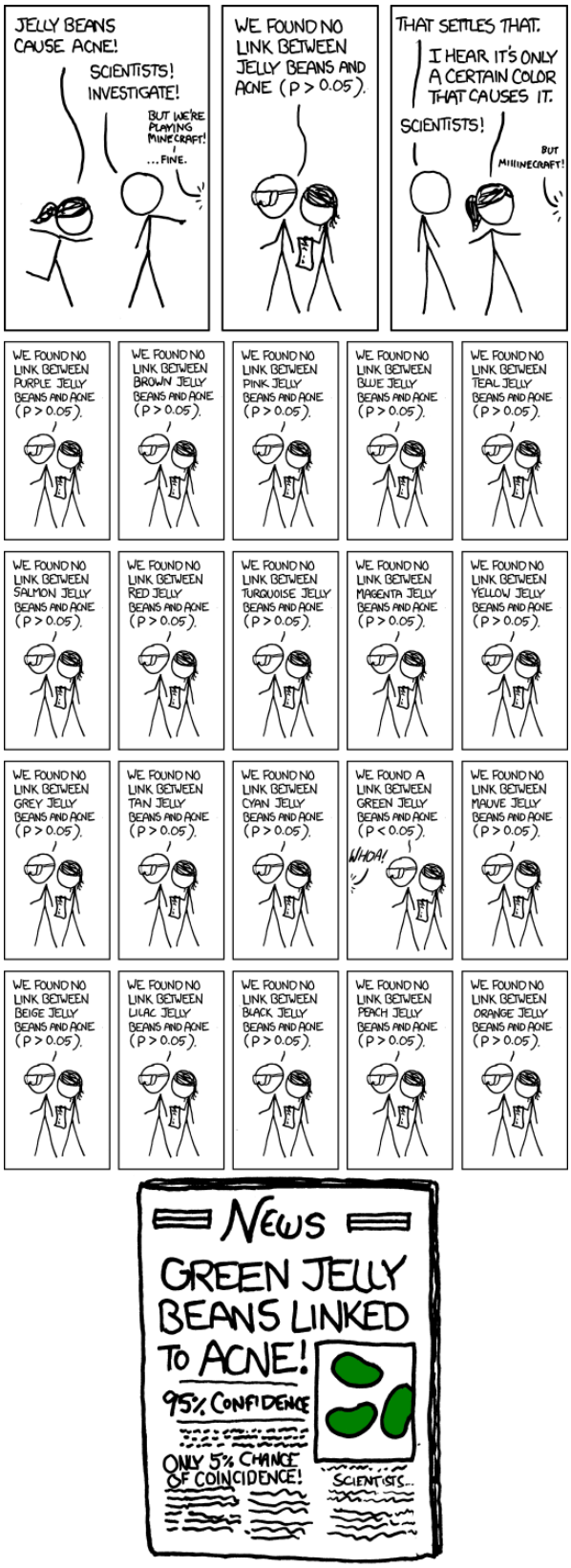
Compute a 95% confidence interval for the mean diameter of ball bearings using the data mentioned in Example 3 (using t-procedures). Does the confidence interval agree with the conclusion of the test conducted in Example 4?

In many situations we don't want to conduct just one statistical test, but many. When we do so, the probability of making a Type I error in *any* test increases.

Suppose, for example, that we perform K tests that are *independent* of each other (a strong and likely incorrect assumption). The following calculations show the probability of making a Type I error in the *study*:

The problem is explained well by Munroe (2011) in the comic xkcd.

One approach to this problem is to adjust the significance levels of the tests to achieve a study Type I error rate. For example, we could work with the above expression to find an appropriate α for each test.


JELLY BEANS CAUSE ACNE!
SCIENTISTS! INVESTIGATE!
BUT WE'RE PLAYING MINECRAFT!
... FINE.
WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE (P > 0.05).
THAT SETTLES THAT.
I HEAR IT'S ONLY A CERTAIN COLOR THAT CAUSES IT.
SCIENTISTS!
BUT MIIIINECRAFT!
WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLUE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN SALMON JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN RED JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TURQUOISE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN MAGENTA JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE (P < 0.05).
WHOA!
WE FOUND NO LINK BETWEEN MAUVE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE (P > 0.05).
NEWS
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!
SCIENTISTS...

The above assumption of independence, though, is strong and unrealistic. Another approach is to use the **Bonferroni inequality**:

This inequality suggests that our α for each test should be:

This may be too strong a correction, though; imagine if we were doing 1000 tests! Thus we don't see this approach used when K is large.

Example 10

A medical researcher tests 1000 genes to see if there is a relationship between gene expression and rate of occurance of cancer. The researcher wants a study Type I error rate of $\alpha = 0.1$. How should we choose α for each test if we assume each test is independent? What if we use the Bonferroni inequality approach?

```
alpha <- 0.1
K <- 1000
1 - (1 - alpha)^(1/K)  # Independence approach
```

```
## [1] 0.000105355
```

```
alpha/K  # Bonferroni approach
```

```
## [1] 1e-04
```

There are other approaches to multiple hypothesis testing. Procedures such as ANOVA and the χ^2 test have the following approach:

1. Execute an overall test to see if any effect is present.
2. If the null hypothesis of no effect is rejected, do a detailed analysis to see where the divergence from this null hypothesis occurs.

Some of the tests discussed in this chapter follow from the **likelihood ratio principle**. The **likelihood ratio statistic** is defined below:

Tests based on the likelihood ratio reject H_0 when the likelihood ratio statistic is "small". The statistic is useful for generating new statistical tests when data follows particular distributions. We can also find more expressive hypotheses using the likelihood ratio.

Example 11

Suppose $X_1, \ldots, X_n$ is an i.i.d. sample, with $X_i \sim \text{Exp}(\mu_i)$. H_0 and H_A are described below:

1. It can be shown that the MLE for $\mu_i = \mu$ when H_0 is true is $\hat{\mu} = \bar{X}$, while the MLE for μ_i in general is $\mu_i = X_i$. Find the corresponding likelihood ratio.

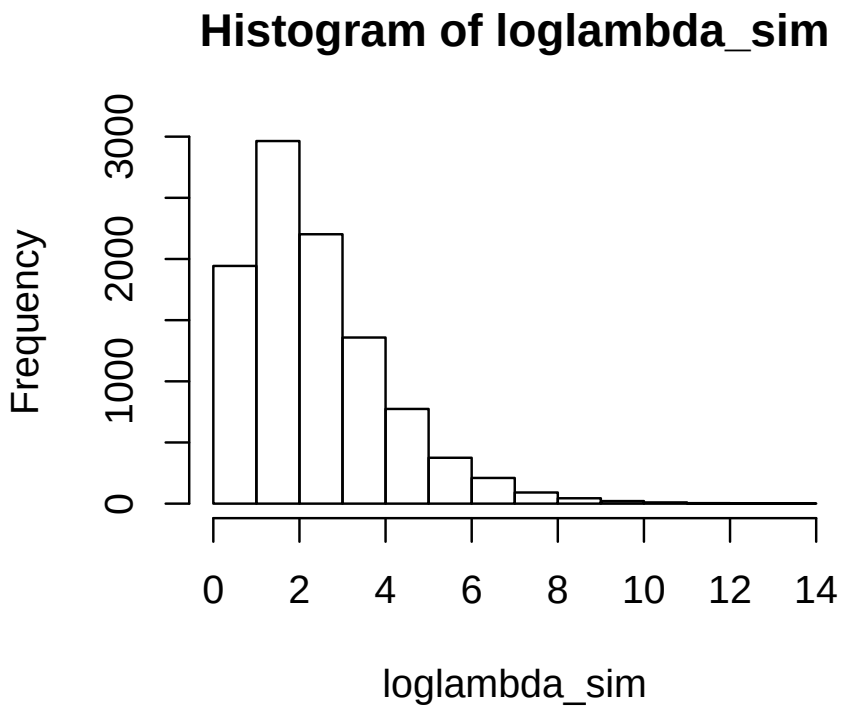
2. Show that if H_0 is true, the distribution of the likelihood ratio does not depend on μ. (This involves showing that μ is a scale parameter.)

3. It is often numerically easier to work with the negative log-likelihood ratio than the likelihood ratio since for large n the likelihood ratio can be very small, and the negative log-likelihood fits with our intuition that large test statistics are evidence against H_0. Write an expression for the negative log-likelihood ratio.

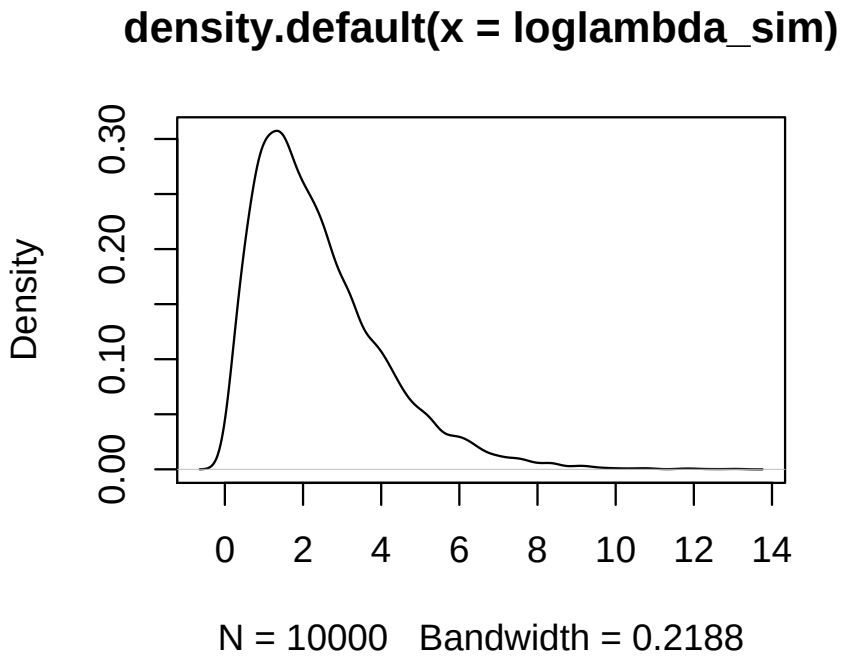
4. Below I estimate the sampling distribution of the negative log-likelihood ratio under H_0 in order to estimate critical values for the negative log-likelihood ratio statistic when $n = 5$:

```
n <- 5
loglambda_sim <- replicate(10000, {
  dat <- rexp(n)  # Can use rate = 1 since the rate does not matter under H0
  -(sum(log(dat)) - n * log(mean(dat)))  # Simulated test statistics
```

```
})
hist(loglambda_sim)
```




```
plot(density(loglambda_sim))
```



```
quantile(loglambda_sim, c(.9, .95, .99, .995, .999, .9995))
```

```
##       90%       95%       99%     99.5%
##  4.597638  5.595884  7.707713  8.505277
##     99.9%    99.95%
## 10.769724 11.840371
```

Suppose now that we are tracking the time between eruptions of a geyeser and we want to know whether the number of eruptions can be modelled with a Poisson process. If that is the case, the time between eruptions is i.i.d. and follow from an exponential distribution with some mean. All we wish to know is whether i.i.d. exponential time is an appropriate model for the time between eruptions (we don't necessarily care about the parameters of the model).

We watch our geyeser and observe the following times (in hours) between eruptions:

```
erupt_time <- c(1, 0.1, 27.6, 6.5, 16.3)
```

Test the appropriate hypotheses and estimate p_{val}. What's the conclusion of the test at a significance level of $\alpha = 0.01$?

```
(teststat <- -(sum(log(erupt_time)) -
                 length(erupt_time) * log(mean(erupt_time))))  # Test statistic
```

```
## [1] 5.982522
```

```
mean(loglambda_sim > teststat)  # Estimated p-value
```

```
## [1] 0.0385
```

```
mean(loglambda_sim > teststat) < 0.01
```

```
## [1] FALSE
```

Methods based on likelihood ratios make strong assumptions about the distribution of the data, specifying the distribution the data takes save for information about the values of some of the parameters. These methods are known as **parametric methods** since they are ultimately probing about the value of parameters of some assumed distribution. **Distribution-free** methods, also known as **non-parametric methods**, make fewer assumptions about the distribution of the data. These methods will not be considered in this course.

Chapter 9: Inferences Based on Two Samples

Introduction

Researchers' questions often address not just one population but two. Frequently the researcher's question doesn't specify a value for a single parameter but gives a relationship between two parameters from two groups so that a relationship can be inferred. For example, while the effect of a new blood pressure drug on blood pressure is good to know, a more interesting question may ask whether a new drug reduces blood pressure more than existing methods.

This chapter discusses procedures intended to compare two samples from two different populations. We will see both how to conduct statistical tests and confidence intervals. The framework for confidence intervals and hypothesis testing hasn't changed. That means that for this chapter a few formulas appropriate for certain contexts are all we need; we don't need to reintroduce the theory.

Section 1: z Tests and Confidence Intervals for a Difference Between Two Population Means

Throughout this chapter I will assume that, unless otherwise stated, we have two different samples, $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ drawn from two independent[106] samples, and that the data within a sample is i.i.d. Let $\mathbb{E}[X_1] = \mu_X$, $\mathbb{E}[Y_1] = \mu_Y$, $\text{Var}(X_1) = \sigma_X^2$, and $\text{Var}(Y_1) = \sigma_Y^2$.

We're often interested in $\Delta = \mu_X - \mu_Y$. What is an estimator for Δ? Is it unbiased?

[106] In Section 3, the samples are not independent, and $m = n$.

What is the variance of this estimator?

Based on this we can find the sampling distribution for $\hat{\Delta}$ that is at least approximately correct when sample sizes are large.

Assume that σ_X^2 and σ_Y^2 are known. Using the above distribution of $\hat{\Delta}$ we can obtain a confidence interval for the true Δ that is appropriate at least approximately for large m and n.[107]

[107] Here we will say that m and n are "large" when both quantities are greater than 40.

When planning a study, if we decide in advance to set $m = n$, we could obtain a formula for n (and thus m as well) that will guarantee a chosen margin of error.

For hypothesis testing we want to make a statement about the value of Δ. The ingredients of this statistical test are listed below.[108]

[108] While the test described below is appropriate for any proposed difference Δ_0, the case $\Delta_0 = 0$ is certainly the most interesting and more frequently seen, as this corresponds to the null hypothesis $H_0 : \mu_X = \mu_Y$; in other words the test determines whether the means of the two populations are the same or differ in some way.

Type II error analysis formulas are provided below.

If we require that both samples have a common sample size we can also obtain a formula that gives sample sizes that produce a test with a specified Type II error rate for a particular Δ_A for a test of significance level α.

Example 1

A tutoring service claims to help understand difficult statistics concepts. You decide to test this. You randomly assign 100 students taking a statistics class to sign up for the tutoring students, while the rest only attend lectures and office hours while learning statistics. At first an equal number of students were assigned to both groups, but after students dropped out, there were 45 students who did not use the tutoring service and 51 who did.[109]

At the end of the course the final exam scores of students from both groups were compared. The students who got tutoring (denoted $X_1, \ldots, X_{51}$) had an average score of 78.79 points. For those who did not get tutoring (denoted $Y_1, \ldots, Y_{45}$), the mean score was 71.09. Assume that $\sigma_X = \sigma_Y = 15$.

[109] This is known as **dropout bias**; if the propensity to drop out does not depend on whether someone belongs to the control or treatment group, there is no problem, but if there is a relationship the results of a study could be biased. This should be accounted for, but we will ignore the problem.

1. Compute a 95% confidence interval for the mean difference in scores. Based on this confidence interval, is there good evidence that the tutoring service improves students' performance on exams?

2. The tutoring service wants the margin of error produced by your study to not exceed 3 points for the exam; this will help the service determine if their product improves students' performance by a letter grade. What sample sizes could achieve this margin of error (while preserving the confidence level)?

3. Perform a statistical test to determine if the tutoring service improves students' scores on exams.

4. The tutoring service wants a statistical test that detects a difference of three points with a Type II error rate of 10% for a test with a Type I error rate of $\alpha = 0.05$. Assuming equal sample sizes for both samples, find sample sizes that leads to a test meeting these requirements.

5. Suppose a statistical has sample sizes of $m = n = 450$, find the Type II error rate when the true difference between the two groups of students is one point.

```r
xbar <- 78.79
ybar <- 71.09

sigma_X <- 15
sigma_Y <- sigma_X

m <- 51
n <- 45

alpha <- .05
(z <- qnorm(alpha/2, lower.tail = FALSE))
```

```
## [1] 1.959964
```

```r
## Part 1
(se <- sqrt(sigma_X^2/m + sigma_Y^2/n))
```

```
## [1] 3.06786
```

```r
(moe <- z * se)
```

```
## [1] 6.012895
```

```r
(est <- xbar - ybar)
```

```
## [1] 7.7
```

```r
c(est - moe, est + moe)
```

```
## [1]  1.687105 13.712895
```

```r
## Part 2
ceiling(z^2 * (sigma_X^2 + sigma_Y^2) / 3^2)
```

```
## [1] 193
```

```r
## Part 3
(z_stat <- (est - 0)/se)
```

```
## [1] 2.509893
```

```r
pnorm(z_stat, lower.tail = FALSE)  # p-value
```

```
## [1] 0.006038389
```

```r
## Part 4
ceiling(((sigma_X^2 + sigma_Y^2) *
    (qnorm(.05, lower.tail = FALSE) + qnorm(.1, lower.tail = FALSE))^2/3^2))
```

```
## [1] 429
```

```
## Part 5
pnorm(qnorm(.05, lower.tail = FALSE) - 1/
        (sqrt(15^2/450 + 15^2/450)))
```

```
## [1] 0.740489
```

Of course we very rarely know what σ_X and σ_Y are and often need to estimate them from the sample. Then we have an estimated standard error

We would use this for our confidence intervals:

In hypothesis testing our test statistic would be

All the rest is the same. Procedures using these CIs and statistics are appropriate for large sample sizes.

Example 2

The sample standard deviation for the students who got tutoring was 14.52 points. The sample standard deviation for the students who did not get tutoring was 11.87 points. Recompute the confidence interval, test statistic, and p_{val} computed in Example 1.

```
(se_est <- sqrt(14.52^2/m + 11.87^2/n))
```

```
## [1] 2.695361
```

```
c(est - z * se_est, est + z * se_est)
```

```
## [1]  2.417189 12.982811
```

```
(z_stat2 <- est/se_est)
```

```
## [1] 2.85676
```

```
pnorm(z_stat2, lower.tail = FALSE)
```

```
## [1] 0.002139946
```

In what contexts can we claim we observe a causal effect in a study? This depends on how the data was generated. If the data was obtained as-is, without being assigned to their groups by the investigators, we may call the study **observational**. If we assigned individuals to groups after the individuals generated their data, we may call the study a **retrospective** observational study. On the other hand, if the investigator assigned individuals randomly to two groups and applied a different treatment depending on group assignment, measuring outcomes after the treatment was applied, we would call the study a **randomized controlled experiment**. The latter type of study allows us to make conclusions about causality, unlike the former.

Section 2: The Two-Sample t Test and Confidence Interval

The procedures from the previous section are appropriate for large sample sizes. When we don't have large sample sizes and we assume the data was drawn from Normal distributions, we can use t procedures.

Suppose we assume $\sigma_X = \sigma_Y$. In most cases this assumption is unrealistic, though there are contexts where the assumption makes sense; for instance, we may be attempting to determine not just the difference in mean but whether two samples come from the same population (and thus would have the same population standard deviation). Then the standard error of $\bar{X} - \bar{Y}$ would be

This is estimated with

Consider now the random variable

This random variable follows a t distribution with $m + n - 2$ degrees of freedom. Knowing this, we can find a confidence interval based on this random variable. Procedures that assume that the two samples have the same standard deviation are known as pooled t procedures.

We could also perform a statistical test.

Example 3

Below are two datasets, with each dataset coming from some distribution. Did the same distribution generate these datasets?

```
x <- c( 7.07, -0.01,  8.30,  5.70,  5.06,
        1.85,  0.74,  2.11, -0.93, 15.88)
y <- c( 1.69,  8.83,  9.11, -1.32,  3.97,
        9.40,  7.60,  4.78,  5.13,  6.38)
```

```
mean(x)
```

```
## [1] 4.577
```

```
sd(x)
```

```
## [1] 5.043597
```

```
mean(y)
```

```
## [1] 5.557
```

```
sd(y)
```

```
## [1] 3.472233
```

1. Find a 90% confidence interval for the population mean, using the pooled t procedure.

2. Using the pooled t test, test whether the datasets have the same distribution or not, at significance level $\alpha = 0.1$.

```r
t.test(x, y, var.equal = TRUE, conf.level = .9)
```

```
##
##  Two Sample t-test
##
## data:  x and y
## t = -0.50611, df = 18, p-value = 0.6189
## alternative hypothesis: true difference in means is not equal to 0
## 90 percent confidence interval:
##  -4.337743  2.377743
## sample estimates:
## mean of x mean of y
##     4.577     5.557
```

The equal variance assumption made by the pooled test, though, is unrealistic. What if we don't make that assumption? Then we can create procedures based on the quantity

This random variable follows approximately a $t(\nu)$ distribution. The formula for ν is given below:

From this random variable we can derive a CI:

Below is a test statistic for a test based on this random variable:

In most situations you should use the test that does not assume equal variances. When this assumption is true, there is a minor gain in power resulting from using the pooled version of the test, but even a slight deviation from this assumption can result in statistics that behave very poorly; the pooled statistic[110] is not at all robust to the assumption of equal variances.

[110] The pooled two-sample test can be viewed as an instance of the likelihood ratio test, while the other test, known as the Wald two-sample t-test, cannot be derived from the likelihood ratio test.

Example 4

Two teams of competitive rowers locked in a bitter rivalry want to know which team is fastest. Instead of a single race, the two teams each engage in 10 time trials along the same 500m river length. The mean time for team 1 (in minutes) is 1.49 with a standard deviation of 0.12, while for team 2 the mean time was 1.37 with a standard deviation of 0.10.

1. Construct a 95% confidence interval for the difference in rowing times.

2. Perform a statistical test to determine whether team 2 is faster than team 1 or not at significance level $\alpha = 0.1$.

```r
x1bar <- 1.49
x2bar <- 1.37

sd1 <- 0.12
sd2 <- 0.10

n <- 10
m <- n

(se1 <- sd1/sqrt(m))
```

```
## [1] 0.03794733
```

```r
(se2 <- sd2/sqrt(n))
```

```
## [1] 0.03162278
```

```r
(std_err_diff <- sqrt(se1^2 + se2^2))
```

```
## [1] 0.04939636
```

```r
(nu <- (se1^2 + se2^2)^2/(se1^4/(m - 1) + se2^4/(n - 1)))
```

```
## [1] 17.43311
```

```r
## Part 1
(tstar <- qt(.975, df = nu))
```

```
## [1] 2.10583
```

```r
c(x1bar - x2bar - tstar * std_err_diff, x1bar - x2bar + tstar * std_err_diff)
```

```
## [1] 0.01597968 0.22402032
```

```r
## Part 2
(test_stat <- (x1bar - x2bar)/std_err_diff)
```

```
## [1] 2.429329
```

```r
pt(test_stat, df = nu, lower.tail = FALSE)
```

```
## [1] 0.01309898
```

Section 3: Analysis of Paired Data

Up until now we have required that $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ be two *independent* samples. However, in many experimental designs, the datasets may not be independent; instead, they may be *paired*. That is, we can view the datasets as $(X_1, Y_1), \ldots, (X_n, Y_n)$.

Examples of independent sample studies and paired sample studies are listed below:[111]

[111] I list these to avoid a common situation in tests: students confusing paired-sample and independent-sample procedures. Don't be another statistic; *know the difference between these tests!*

We are still primarily interested in $\Delta = \mu_D = \mu_X - \mu_Y$, but since the data is paired, we don't approach inference in the same way. Instead of treating $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$ separately, we work with a sample of *differences*:

When we do this, comparing two populations reduces to the one-sample case we saw in chapter 8. Below are CIs and statistical tests in this context:[112]

[112] What happens when we use the two independent sample procedures in the presence of paired data? The biggest difference is that the variance of our estimator for Δ is no longer correct, since the true variance is

$$\text{Var}\,(\bar{X} - \bar{Y}) = \frac{\sigma_X^2 + \sigma_Y^2 - 2\rho\sigma_X\sigma_Y}{n}$$

For independent samples, $\rho = 0$, but that's likely not the case for paired data; in fact, usually $\rho > 0$. As a result

---

Example 5

A drug manufacturer wants to determine if a new weight loss supplement leads to weight loss in subjects. To determine if the supplement leads to weight loss, the manufacturer selects a cohort of six participants to participate. The subjects' weights are measured prior to taking the supplement, then after two months the subjects' weights will be measured again. Below are subjects' weights both before and after taking the supplement:

```
(supp_weight_loss <- data.frame(
  "before" = c(221, 139, 253, 230, 186, 161),
  "after"  = c(209, 121, 230, 220, 182, 162)
))
```

```
##   before after
## 1    221   209
## 2    139   121
## 3    253   230
## 4    230   220
## 5    186   182
## 6    161   162
```

1. Compute the dataset of differences, D_i.

2. Construct a 90% confidence interval for the mean difference in weight loss.

3. Conduct a hypothesis test to determine whether the supplement leads to weight loss. Use a significance level of $\alpha = 0.1$ to decide whether there is a statistically significant difference in weight after taking the supplement.

```
## Compute CI
with(supp_weight_loss,
     t.test(before, after, conf.level = .9, paired = TRUE,
            alternative = "two.sided")
)
```

```
##
##  Paired t-test
##
## data:  before and after
## t = 3.0587, df = 5, p-value = 0.02814
## alternative hypothesis: true difference in means is not equal to 0
## 90 percent confidence interval:
##   3.753293 18.246707
## sample estimates:
## mean of the differences
##                      11
```

```
## Statistical test
with(supp_weight_loss,
     t.test(before, after, paired = TRUE, alternative = "greater")
)
```

```
##
##  Paired t-test
##
## data:  before and after
## t = 3.0587, df = 5, p-value = 0.01407
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  3.753293      Inf
## sample estimates:
## mean of the differences
##                      11
```

Prior to conducting a study, should you opt for an independent-sample study or a paired-sample study? Below are pros and cons of the two:

In general, if there is a lot of variability within a population and a large correlation resulting from pairing, then a paired sample may be preferred to independent samples, while if there is not a lot of variability and the sample is not large, we may prefer an independent-sample procedure.

Section 4: Inferences Concerning a Difference Between Population Proportions

Suppose that instead of being interesterested in the difference between two means of quantitative variables we are interested in the difference in the population proportions from two different populations. For example, we may want to compare the rate at which a disease appears in men versus women, or compare political affiliation across different demographic groups. In any case, there is one population at which the probability of a "success" is p_X and another population where the probability of a "success" is p_Y. If we were conducting a hypothesis test, we may want to see if $p_X = p_Y$ or not.

We say that we have two independent samples of i.i.d. binomial data, $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$. We're interested in estimating $p_X - p_Y$. The natural estimator for this parameter is:

The variance and standard error of this estimator is

When m and n are large, the following random variable follows an approximately standard Normal distribution:

We can use this to construct confidence intervals and statistical tests.

Let's discuss statistical testing first. When H_0 is true, $p_X = p_Y = p$. We need to estimate p, and can do so using the pooled sample (the sample that includes both datasets, $X_1, \ldots, X_m, Y_1, \ldots, Y_n$). The resulting estimator is

We then have the following test, appropriate for large sample sizes:

Type II errors don't depend on $p_X - p_Y$ but instead on each specific p_X and p_Y, so the Type II error function is denoted with $\beta(p_X, p_Y)$.

The following formulas can be used for Type II error analysis:

When conducting sample size planning, if we set $m = n$ and want to detect a difference in proportions of $p_X - p_Y = d$, after guessing p_X and p_Y we get a guessed sample size of

This is the appropriate sample size for a one-sided alternative

hypothesis; for a two-sided alternative, replace α with $\alpha/2$.

Example 6

A pharmaceutical company plans to release a new vaccine intended to reduce the risk of contracting the influenza virus. The company plans to test the vaccine by randomly assigning study participants to a control group and a treatment group. Individuals in the treatment group will receive the new vaccine, while individuals in the control group will receive no treatment.[113] The experiment is conducted in a double-blind fashion; that is, neither patients nor experimental staff will know which patient received which vaccine until after the experiment is complete.

[113] This is ethically suspect, but ignore ethics for now.

1. The experimenters plan on assigning an equal number of subjects to both control and treatment groups. They want to be able to detect a 5% difference in contraction rate (in the new vaccine's favor) 95% of the time. The current influenza contraction rate is believed to be 20%. The planned significance level is $\alpha = 0.1$. Based on this, what sample size should be used?

2. Using the answer from above, what is the probability of making a

Type II error when the new vaccine reduces the rate of influenza contraction by only 1%?

3. When the experiment was actually conducted, after participants left the study for various reasons, the number of individuals in the control group was 886 and the number of individuals in the treatment group was 890. 183 individuals in the control group contracted the flu, while 175 contracted the virus in the treatment group. Test whether the vaccine reduced the occurance of influenza. What is the conclusion?

```
p_X <- .2
d <- .05
alpha <- .1
beta <- .05

## Part 1
(n <- ceiling((qnorm(alpha, lower.tail = FALSE) *
                sqrt(p_X + (p_X - d) * ((1 - p_X) + (1 - (p_X - d)))/2) +
                qnorm(beta, lower.tail = FALSE) * sqrt(p_X * (1 - p_X) +
                                                         (1 - p_X) *
                                                         (1 - (p_X - d)
                                                          )))/d^2))

## [1] 895

m <- n

## Part 2
(sigma <- sqrt(p_X * (1 - p_X)/m + (p_X - d) * (1 - (p_X - d))/n))

## [1] 0.01792286

(pbar <- (m * p_X + n * (p_X - d))/(m + n))

## [1] 0.175

(qbar <- (m * (1 - p_X) + n * (1 - (p_X - d)))/(m + n))

## [1] 0.825

pnorm((qnorm(alpha, lower.tail = FALSE) * sqrt(pbar * qbar * (1/m + 1/n)) - d)/
        sigma)

## [1] 0.06611087

## Part 3
(phat <- (183 + 175)/(886 + 890))

## [1] 0.2015766

(z <- (183/886 - 175/890)/sqrt(phat * (1 - phat) * (1/886 + 1/890)))

## [1] 0.5208789

pnorm(z, lower.tail = FALSE)  # p-value

## [1] 0.3012256
```

We can construct a large-sample confidence interval for the difference in proportions using the formula:[114]

[114] This interval should be appropriate when $m\hat{p}_X$, $n\hat{p}_Y$, $m(1-\hat{p}_X)$, and $n(1-\hat{p}_X)$ are all at least 10.

Example 7

Construct a 90% CI for the difference in influenza contraction rates based on the data in Example 6. Does this CI agree with the conclusion of the test?[115]

[115] Looking at the formulas for the test statistic and the CI, we should not believe that the CI will necessarily agree with the test.

```
(phat_X <- 183/886)
```

```
## [1] 0.2065463
```

```
(phat_Y <- 175/890)
```

```
## [1] 0.1966292
```

```
m <- 886
n <- 890
```

```
(se <- qnorm(.05, lower.tail = FALSE) * sqrt(phat_X * (1 - phat_X)/m +
                                               phat_Y * (1 - phat_Y)/n))
```

```
## [1] 0.03131543
```

```
c("Lower" = phat_X - phat_Y - se, "Upper" = phat_X - phat_Y + se)
```

```
##       Lower       Upper
## -0.02139837  0.04123249
```

Section 5: Inferences Concerning Two Population Variances

So far we have been interested in comparing μ_X and μ_Y or p_X and p_Y. Sometimes, though, we may be interested in comparing σ_X^2 and σ_Y^2.

Let $N \sim \chi^2(\nu_n)$ and $D \sim \chi^2(\nu_d)$. Consider the random variable

This random variable follows the $F(\nu_n, \nu_d)$ distribution. The density curve for this distribution is illustrated below:

The pdf and cdf of the $F(\nu_n, \nu_d)$ distribution is difficult to describe but I list the expected value and variance of this distribution below:

Suppose $F \sim F(\nu_n, \nu_d)$. Let f_{α,ν_n,ν_d} satisfy $\mathbb{P}\left(F \geq f_{\alpha,\nu_n,\nu_d}\right) = \alpha$. We will call f_{α,ν_n,ν_d} a critical value of the $F(\nu_n, \nu_d)$ distribution. Critical values (and thus some values of the cdf of) the $F(\nu_n, \nu_d)$ distribution are listed in Table A.9, for select ν_n and ν_d. ν_n is called the **numerator degrees of freedom** and ν_d is called the **denominator degrees of freedom**. An important identity for critical values of the $F(\nu_n, \nu_d)$ distribution is

---

Example 8

Let $F \sim F(4, 9)$.

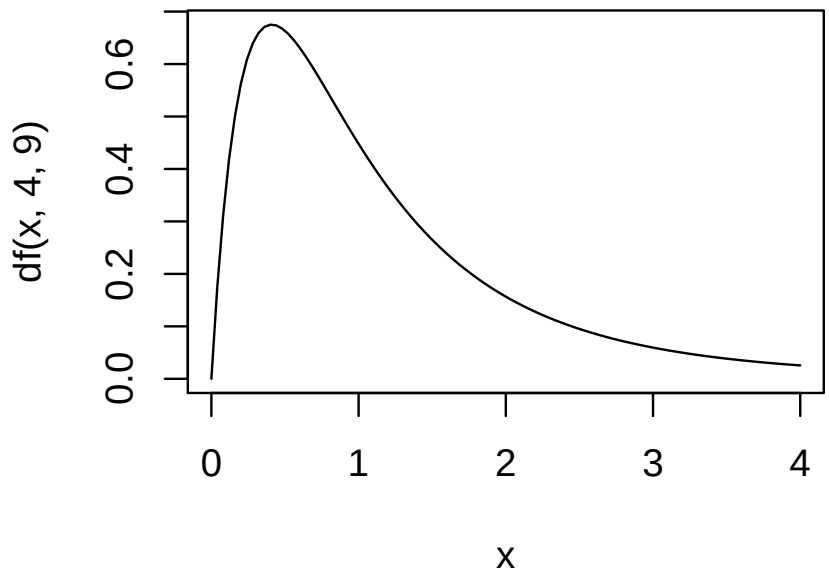
1. Compute $\mathbb{E}[F]$ and $\text{Var}(F)$.

2. Compute $\mathbb{P}(F \leq 3.63)$.

3. Find $f_{.01,4,9}$.

4. Find $f_{.999,4,9}$

```r
curve(df(x, 4, 9), 0, 4)
```




```r
## Part 1
(mu_f <- integrate(function(x) {x * df(x, 4, 9)}, 0, Inf))
```

```
## 1.285714 with absolute error < 7.8e-06
```

```r
(var_f <- integrate(function(x) {(x - mu_f$value)^2 * df(x, 4, 9)}, 0, Inf))
```

```
## 1.818367 with absolute error < 5.5e-05
```

```r
## Part 2
pf(3.63, 4, 9)
```

```
## [1] 0.9498937
```

```r
## Part 3
qf(.01, 4, 9, lower.tail = FALSE)
```

```
## [1] 6.422085
```

```r
## Part 4
qf(.999, 4, 9, lower.tail = FALSE)
```

```
## [1] 0.0206294
```

The F distribution matters because when $X_1, \ldots, X_m$ is an i.i.d. sample with $X_1 \sim N(\mu_X, \sigma_X)$ and $Y_1, \ldots, Y_n$ is an i.i.d. sample with $Y_1 \sim N(\mu_Y, \sigma_Y)$, if S_X^2 is the sample variance for the first dataset and S_Y^2 is the sample variance for the second dataset, we can find a distribution for S_X^2 / S_Y^2.

This distribution can be used for deriving confidence intervals and statistical tests for σ_X^2/σ_Y^2 and thus make statements about the relationship between σ_X^2 and σ_Y^2.[116]

Below I describe a hypothesis test for checking the relationship between σ_X^2 and σ_Y^2.

[116] Thus we also have statements for σ_X and σ_Y's relationship when we take square roots appropriately.

We can also derive formulas for the confidence interval for σ_X^2/σ_Y^2.

Example 9

The standard deviation of the returns of a stock is called the stock's volatility in finance. Two stocks, CGM and UOU, are believed to have Normally distributed returns. Some returns of these stocks are listed below:

```
cgm <- c(-0.004,  0.006,  0.002, -0.023, -0.006, -0.004,  0.023, -0.011,
          0.001,  0    , -0.004)
uou <- c( 0,     -0.011,  0.015,  0.005, -0.012,  0.003, -0.009,  0.005)
```

```
format(var(cgm), scientific = FALSE)
```

```
## [1] "0.0001267636"
```

```
format(var(uou), scientific = FALSE)
```

```
## [1] "0.00008971429"
```

1. Find a 90% confidence interval for $\sigma_{\text{CGM}}/\sigma_{\text{UOU}}$. Based on this CI, is it plausible that the two stocks have the same volatility?

2. Perform a statistical test to check whether the two stocks have the same volatility. Does the result of the test agree with the confidence interval's conclusion?

```
(res <- var.test(cgm, uou, conf.level = .9))
```

```
##
##  F test to compare two variances
##
## data:  cgm and uou
## F = 1.413, num df = 10, denom df = 7,
## p-value = 0.6643
## alternative hypothesis: true ratio of variances is not equal to 1
## 90 percent confidence interval:
##  0.3885498 4.4303192
## sample estimates:
## ratio of variances
##            1.41297
```

```
sqrt(res$conf.int)  # Need to take square root to get CI of volatility ratio
```

```
## [1] 0.6233377 2.1048323
## attr(,"conf.level")
## [1] 0.9
```

Bibliography

Aschwanden, C. (2015). Science isn't broken.

Aschwanden, C. (2016). Failure is moving science forward.

Bernoulli, D. (1954). Exposition of a new theory on the measurement of risk. *Econometrica*, 22(1):23–36.

Boos, D. D. and Stefanski, L. A. (2011). P-value precision and reproducibility. *The American Statistician*, 65(4):213–221.

Box, J. F. (1987). Guinness, gosset, fisher, and small samples. *Statistical Science*, 2(1):45–52.

Buja, A., Hare, E., and Hofmann, H. (2015). *discreteRV: Create and Manipulate Discrete Random Variables*. R package version 1.2.2.

Devore, J. (2015). *Probability and Statistics for Engineering and the Sciences*. Cengage Learning.

George, G. (2004). Testing for the independence of three events. *Mathematical Gazette*, 88.

Hahn, G. J. and Meeker, W. Q. (2011). *Statistical intervals: a guide for practitioners*, volume 92. John Wiley & Sons.

Husak, G. J., Michaelsen, J., and Funk, C. (2007). Use of the gamma distribution to represent monthly rainfall in Africa for drought monitoring applications. *International Journal of Climatology*, 27(7):935–944.

Macdonell, W. R. (1902). On criminal anthropometry and the identification of criminals. *Biometrika*, 1(2):177–227.

Maltamo, M., Puumalainen, J., and Pivinen, R. (2007). Comparison of beta and Weibull functions for modelling basal area diameter distribution in stands of *pinus sylvestris* and *picea abies*. *Scandinavian Journal of Forest Research*, 10(1-4):284–295.

Mandelbrot, B. and Hudson, R. L. (2007). *The Misbehavior of Markets: A fractal view of financial turbulence*. Basic books.

Munroe, R. (2011). Significant.

Salsburg, D. (2002). *The Lady Tasting Tea: How Statistics Revolutionized Science in the Twentieth Century*. Henry Holt and Company.

Slutsky, E. (1925). Über stochastische Asymptoten und Grenzwerte.

Stevens, S. S. (1946). On the Theory of Scales of Measurement. *Science*, 103:677–680.

Verzani, J. (2014). *Using R for Introductory Statistics, Second Edition*. Chapman & Hall/CRC The R Series. Taylor & Francis.

Wasserstein, R. L. and Lazar, N. A. (2016). The ASA's statement on p-values: Context, process, and purpose. *The American Statistician*, 70(2):129–133.

Wikipedia (2018). Binomial proportion confidence interval — Wikipedia, the free encyclopedia. `http://en.wikipedia.org/w/index.php?title=Binomial%20proportion%20confidence%20interval&oldid=832472232`. [Online; accessed 06-April-2018].

Wikipedia (2018). Unbiased estimation of standard deviation — Wikipedia, the free encyclopedia. [Online; accessed 12-June-2018].